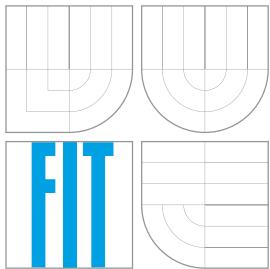


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

FUNDAMENTÁLNÍ ANALÝZA NUMERICKÝCH DAT PRO AUTOMATICKÝ TRADING

FUNDAMENTAL ANALYSIS OF NUMERICAL DATA FOR AUTOMATIC TRADING

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. PETR HUF

VEDOUCÍ PRÁCE
SUPERVISOR

doc. Dr. Ing. JAN ČERNOCKÝ

BRNO 2016

Zadání diplomové práce

Řešitel: **Huf Petr, Bc.**

Obor: Počítačová grafika a multimédia

Téma: **Fundamentální analýza numerických dat pro automatický trading**
Fundamental Analysis of Numerical Data for Automatic Trading

Kategorie: Umělá inteligence

Pokyny:

1. Seznamte se s literaturou na téma využití fundamentální informace pro automatický trading
2. Navrhněte, kde získat data, která potenciálně ovlivňují finanční trhy (např. počasí, politické preference, ...) a sestavte nástroje pro jejich získání a čištění.
3. Navrhněte množinu cílových finančních produktů.
4. Navrhněte systém pro predikci budoucích kursů těchto produktů na bázi strojového učení.
5. Otestujte na zvolené datové sadě a diskutujte výsledky.
6. Navrhněte (technicky i ekonomicky) interfacování s reálnými obchodními systémy.
7. Vytvořte krátké video a/nebo poster presentující Vaši práci.

Literatura:

- podle doporučení vedoucího práce

Při obhajobě semestrální části projektu je požadováno:

- Body 1-3, částečné rozpracování 4.

Podrobné závazné pokyny pro vypracování diplomové práce naleznete na adrese
<http://www.fit.vutbr.cz/info/szz/>

Technická zpráva diplomové práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap, které byly vyřešeny v rámci dřívějších projektů (30 až 40% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Černocký Jan, doc. Dr. Ing., UPGM FIT VUT**

Datum zadání: 1. listopadu 2015

Datum odevzdání: 25. května 2016

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií
Ústav počítačové grafiky a multimédií
612 66 Brno, Božetěchova 2

doc. Dr. Ing. Jan Černocký
vedoucí ústavu

Abstrakt

Tato práce se zabývá využitím fundamentální analýzy v automatickém obchodování. Technická analýza využívá k predikci ceny hlavně její historické hodnoty a indikátory z této ceny odvozené. Fundamentální analýza naopak využívá informace z různých zdrojů k predikci cenového signálu, přičemž v této práci byly zkoumány pouze kvantitativní zdroje dat. Konkrétně se jedná o počasí, Forex, Google Trends, WikiTrends, historické ceny různých futures a souhrnná fundamentální data (porodnost, migrace, ...). Takto získána data jsou zpracovávána LSTM neuronovou sítí, která provádí predikci ceny vybraných akcií. Na základě této predikce je postaven obchodní systém. Experimenty v této práci ukazují na zlepšení výsledků obchodního systému až o 8% v úspěšnosti predikce díky zapojení fundamentální analýzy.

Abstract

This thesis is aimed to exploitation of fundamental analysis in automatic trading. Technical analysis uses historical prices and indicators derived from price for price prediction. On the opposite, fundamental analysis uses various information resources for price prediction. In this thesis, only quantitative data are used. These data sources are namely weather, Forex, Google Trends, WikiTrends, historical prices of futures and some fundamental data (birth rate, migration, ...). These data are processed with LSTM neural network, which predicts stocks prices of selected companies. This prediction is basis for created trading system. Experiments show major improvement in results of the trading system; 8% increase in success prediction accuracy thanks to involvement of fundamental analysis.

Klíčová slova

Obchodování, neuronové sítě, data, počasí, fundamentální analýza, akcie, burza, strojové učení, rekurentní neuronové sítě

Keywords

Trading, neural networks, data, weather, fundamental analysis, shares, stock exchange, machine learning, recurrent neural networks

Citace

HUF, Petr. *Fundamentální analýza numerických dat pro automatický trading*. Brno, 2016. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Černocký Jan.

Fundamentální analýza numerických dat pro automatický trading

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením pana doc. Dr. Ing. Jana Černockého. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Petr Huf
23. května 2016

Poděkování

Tímo bych chtěl poděkovat svému vedoucímu bakalářské práce, doc. Dr. Ing. Janu Černockému, za odborné vedení práce, poskytnutí mnoha rad a nápadů, lidský přístup a příjemnou atmosféru. Také děkuji své rodině a přítelkyni za podporu a trpělivost při psaní této práce.

© Petr Huf, 2016.

Tato práce vznikla jako školní dílo na Vysokém učení technickém v Brně, Fakultě informačních technologií. Práce je chráněna autorským zákonem a její užití bez udělení oprávnění autorem je nezákonné, s výjimkou zákonem definovaných případů.

Obsah

1	Úvod	4
1.1	Motivace	4
1.2	Existující práce	5
1.3	Struktura práce	6
2	Obchodování na finančních trzích	7
2.1	Princip tvorby ceny	7
2.2	Grafy	7
2.3	Typy produktů	9
2.3.1	Forex	9
2.3.2	Indexy	10
2.3.3	Akcie	11
2.3.4	Komodity	12
2.3.5	Opce	12
2.3.6	Spready	12
2.4	Analýzy	12
2.4.1	Technická analýza	12
2.4.2	Fundamentální analýza	13
3	Neuronové sítě	16
3.1	Základní model neuronové sítě	16
3.1.1	Princip výpočtu	17
3.2	Typy neuronových sítí	20
3.2.1	Použití	21
3.3	LSTM neuronová síť	21
3.4	Dlouhodobá predikce	24
3.5	Dropout	25
4	Data	28
4.1	Výběr produktů	28
4.2	Data produktů	29
4.3	Počasí	29
4.4	Forex	30
4.5	Google Trends	30
4.6	WikiTrends	31
4.7	Futures	31
4.8	Fundamentals	31

5 Předzpracování dat	35
5.1 Čištění	35
5.1.1 Globální čištění	35
5.1.2 Akcie	36
5.1.3 Počasí	36
5.1.4 Forex	37
5.1.5 Futures	37
5.1.6 Ostatní	37
5.2 Chybějící data	37
5.3 NN formát dat	39
5.4 Normalizace	39
5.5 Redukce dimenze	41
6 Návrh systému	43
6.1 Architektura systému	43
6.1.1 Data center	43
6.1.2 Rnn center	44
6.1.3 Trading center	44
7 Implementace systému	45
7.1 Struktura systému	45
7.2 Příprava jednotlivých zdrojů dat	45
7.2.1 Akcie	45
7.2.2 Historické ceny akcií	45
7.2.3 Počasí	46
7.2.4 Forex	46
7.2.5 Google Trends	46
7.2.6 WikiTrends	46
7.2.7 Futures	47
7.2.8 Fundamentals	47
7.3 Knihovny	47
7.3.1 LSTM neuronová síť	48
7.3.2 PCA	48
7.3.3 Multiple Imputation	48
7.3.4 Maticové operace	48
7.3.5 Ostatní	48
8 Experimentování s technikami	49
8.1 Akcie – základní konfigurace	51
8.2 Dlouhodobá predikce	52
8.3 Dropout	55
8.4 Transformace dat	56
8.5 Normalizace	58
8.5.1 Z-score	58
8.5.2 Min-max	59
8.6 Chybějící data	61

9 Experimentování s fundamentálními daty	66
9.1 Počasí	66
9.2 Forex	72
9.3 Google Trends	73
9.4 WikiTrends	76
9.5 Futures	78
9.6 Fundamentals	79
9.7 Obchodní systém	81
9.8 Vyhodnocení	84
10 Živé obchodování	87
10.1 Technické provedení	87
10.2 Finanční záležitosti	87
11 Závěr	89
11.1 Možnosti pokračování	90
Literatura	91
Přílohy	96
Seznam příloh	97
A Obsah CD	98
B Manuál	99
B.1 Prekvizity	99
B.2 Konfigurace	99
B.3 Výpočet systému	99
C Plakát	101

Kapitola 1

Úvod

Obchodování je každodenní součástí života téměř každého člověka. Nejčastěji jsou směňovány peníze za zboží a služby. Cena je předem daná a je určena situací na trhu. Trh je označení pro místo, kde jsou obchody prováděny. Situace na trhu je ovlivňována mnoha faktory — množstvím konkurence, množstvím nabízeného zboží/služby, velikostí nabídky a poptávky apod. Na základě těchto faktorů se cena zboží/služby dynamicky mění v průběhu času.

Tato práce se bude zabývat obchodováním na burzovním trhu. Burzovní trh je místo, kde jsou směňovány různé produkty¹. Na obchodování vždy dohlíží nějaká centrální autorita. To má zaručit férové obchodování pro všechny obchodníky. Ti obchodují na burze prostřednictvím brokerů, kteří mají na burzu zajištěný přístup. Za tuto službu si úctují určité poplatky.

Na burze lze obchodovat například s firemními akcemi nebo komoditami. Cena těchto produktů se taktéž v průběhu času mění. Díky tomu lze na pohybu ceny vydělávat. V ideálním případě obchodník nakoupí akcie firmy. Následně dojde k prudkému růstu ceny akcií. Tento růst může být způsoben například změnou majitele firmy, zveřejněním pozitivního výsledku v ekonomickém hospodaření firmy v posledním čtvrtletí, ale třeba i uvalením obchodního embarga na Rusko. Po vzrůstu ceny obchodník akcie firmy prodá za vyšší cenu, než je nakoupil. Rozdíl v ceně za nákup a za prodej je obchodníkův zisk.

Ve výše uvedeném scénáři je naprosto klíčové správné načasování. Obchodník potřebuje vědět, kdy dojde ke vzrůstu ceny a před tímto vzrůstem akcie nakoupit. Po nákupu zase potřebuje vědět, kdy cena už dále růst nebude a akcie prodat, jinak se cena může snížit, a tím se obchodníkův zisk opět zmenší. Jelikož do budoucnosti nikdo nevidí, ani obchodník nemůže nikdy s jistotou vědět, že cena poroste. Proto obchodníci používají nejrůznější metody, které jim mají alespoň napovědět, jak se cena bude v budoucnu chovat. A právě jedna taková metoda je tématem této diplomové práce.

Ze semestrálního projektu byly využity kapitoly 1–4. Tyto kapitoly byly v této práci dále rozepsány.

1.1 Motivace

Dříve se obchodovalo přímo na tzv. *pitu*. Jednalo se o způsob, kdy obchody byly uzavírány osobami, které byly na burze fyzicky přítomny. Broker tam měl často svého zástupce. Obchodník pak telefonem zadával nákupní a prodejní příkazy brokerovi, který příkazy následně

¹Pod produkty rozumíme téměř cokoliv, co se běžně obchoduje – akcie, indexy, měny apod.

posílal na obchodní parket (pit).

S rozšířením počítačů pitové obchodování postupně ustupovalo a bylo nahrazeno obchodováním elektronickým. V tomto případě obchodník používá obchodní software, který nabízí broker. V tomto softwaru zadává obchodní příkazy. Ty jsou pak odesílány přes brokera na burzovní servery, kde jsou příkazy zpracovávány. Vše tedy probíhá elektronicky, bez nutnosti lidského řízení.

Právě s přechodem k elektronickému obchodování a zapojení počítačů do tohoto procesu, dostala velkého rozmachu technická analýza. Jedná se o způsob, kdy se analyzují historické ceny určitého produktu, a z této historie se obchodník snaží určit budoucí směrování cen. K tomu využívá samotné historické ceny, jejich krouzavý průměr, různé úrovně apod.

Kromě technické analýzy se už asi od roku 1999 [5] začalo rozšiřovat vysokofrekvenční obchodování (high frequency trading). Pro tento obchodní styl jsou charakteristické obchody trvající v řádu milisekund nebo sekund. Je vyžadováno rychlé internetové připojení, co nejmenší odezva a rychle rozhodování. Toto rozhodování mají na starosti různé algoritmy a matematické modely. Ty jako vstupy často používají tzv. *order book* [12]. Jedná se o data obsahující všechny objednávky, které jsou na burzu zaslány.

V obou výše zmíněných přístupech se tedy rozhodování provádí na základě historických cen a objednávek. Naproti těmu dvěma přístupům stojí fundamentální analýza. Tento způsob obchodování do rozhodovacího procesu přibírá vlastně cokoliv, co může ovlivnit pohyb ceny. Jako jednoduchý příklad lze uvést třeba výroční zprávy o hospodaření různých firem nebo i politická situace v Číně. Tento typ rozhodování je těžko automatizovatelný, a proto se v automatických obchodních systémech příliš nevyskytuje. To velmi ubírá na jejich popularitě. Kromě toho se totiž tento přístup při manuálním provádění obtížně zpětně testuje.

1.2 Existující práce

Tato práce se tedy snaží jít *proti proudu* a vytvořit automatickou obchodní strategii, která provádí rozhodování na základě fundamentální analýzy. Podobné práce již existují. V předchozím roce vznikly na této škole hned dvě takové diplomové práce. První se zabývá zpracováním textu získaného z příspěvků na Twitteru [18]. Druhá práce zpracovává titulky z finančních zpráv spolu s daty z forexového kalendáře [24]. Na rozdíl od těchto prací, v této práci se budou zpracovávat data pouze kvantitativní.

Velký souhrn prací zabývajících se zpracováním fundamentálních informací pro predikci cen se nachází v [27]. Tyto práce se sice zabývají zpracováním textových informací na rozdíl od kvantitativních, ale jako fundamentální data používají ekonomické zprávy, příspěvky na Twitteru nebo finanční výkazy firem. V této práci ale taková data použita nebudou (viz kapitola 4). Vzhledem k charakteru zpracovávaných dat má k této práci blízko například [36]. Tato práce hledala souvislost mezi daty v Google Trends pro vybrané fráze a cenami vybraných akciových titulů. Poté byla postavena i konkrétní obchodní strategie. Výsledkem této práce bylo zjištění, že tato souvislost opravdu existuje a data z Google Trends jdou použít jako základ obchodní strategie. Proto byla část vybraných frází použita pro Google Trends i v této práci.

Druhou zajímavou prací je [25]. V tomto případě byl zkoumán vztah počtu editací a zobrazení vybraných článku na Wikipedii a akciovým indexem Dow Jones Industrial Average. Opět byla postavena jednoduchá obchodní strategie. Výsledky naznačují, že hledaný vztah existuje. Například při zvýšeném počtu zobrazení vybraných článků dochází následovně k poklesu tohoto akciového indexu. Opravdu účinná byla ale tato obchodní strategie

až v delších časových rámcích (měsíce, roky).

Poslední bude zmíněna práce [15]. V této práci byla zkoumána závislost množství sluňecního svitu a růstem akciového indexu. Tento indikátor byl použit spíše jako podpůrný prostředek pro obchodní strategii. Závislost každopádně byla opět prokázána. Naopak se nepodařilo prokázat souvislost růstu akciového indexu a dešťových nebo sněhových přeháněk.

1.3 Struktura práce

V kapitole 2 je diskutována problematika obchodování obecně. Jsou zde zmíněny nejznámější obchodovatelné produkty a přístupy k obchodovování. V další kapitole 3 je detailně popsán princip neuronových sítí a jejich použití. Nalézá se zde i výčet problémů s neuronovými sítěmi, které přímo souvisí s touto prací. Kromě představení těchto problémů se zde nachází i nástin jejich řešení. Další kapitola 4 obsahuje popis a získání dat použitých v této práci. Jejich čištěním a dalším zpracováním se pak zabývá kapitola 5. Kapitola 6 se zabývá obecným popisem systému, tedy jeho architekturou. Následující kapitola 7 už popisuje přímo jeho implementaci. Hlavně je zde popsán mechanismus získání dat a vypsány externí použité knihovny. Další kapitolou je 8. Ta obsahuje rozsáhlý popis provedených experimentů a jejich komentáře. Tyto experimenty se týkají pouze vybraných technik. Experimenty s fundamentálními daty jsou popsány až v následující kapitole 9. Předposlední kapitola 10 pak popisuje problematiku živého nasazení vyvíjené strategie. Poslední kapitola 11 už obsahuje závěr práce a diskutuje dosažené výsledky. Nakonec jsou zde ještě popsány možnosti dalšího vývoje implementovaného obchodního systému.

Kapitola 2

Obchodování na finančních trzích

Princip obchodování na finančních trzích, a jak pomocí něj dosáhnout zisku, byl nastíněn již v úvodu. V této kapitole tedy budou podrobněji rozepsány vybrané aspekty obchodování, které si zaslouží vyšší pozornost.

2.1 Princip tvorby ceny

To, jak se cena mění (pohyb nahoru nebo dolů), je určeno aktuální nabídkou a poptávkou. Obchodníci totiž zasílají na burzu limitní obchodní příkazy, které říkají, za jakou cenu jsou určitý produkt ochotni prodat nebo koupit. Cena *ask* označuje nejlepší cenu za kterou lze produkt koupit a cena *bid* pak označuje nejlepší cenu za kterou lze produkt prodat. Kromě limitních příkazů obchodníci zasílají také příkazy typu *market*. Jedná se o příkaz, kterým obchodník nakoupí nebo prodá za nejlepší možnou cenu (tedy *ask* nebo *bid*).

Rozdíl mezi hodnotami *ask* a *bid* se nazývá *spread*. Čím je spread vyšší, tím hůře pro obchodníka. Kromě ceny *ask* a *bid* se ještě udává „virtuální“ cena *last*. To je cena, za kterou byl proveden poslední obchod a která se běžně používá pro zobrazování grafů.

To, jak dochází k pohybu ceny, si nyní uvedeme na následujícím příkladu [8]. Určitá komodita se obchoduje na trhu. V tabulce 2.1 lze vidět, že aktuální cena *ask* je 931, *bid* je 930 a *last* je 930 (uvedeno žlutě). Na každé cenové hladině je vyznačen počet kontraktů k prodeji/nákupu za uvedenou cenu (limitní příkazy). Přepokládejme že libovolný obchodník chce ihned (příkaz *market*) prodat 12 kontraktů. V tabulce 2.2 se tedy počet nabízených kontraktů ke koupi snížil ze 44 na 32. Následně jiný obchodník chce nakoupit ihned 55 kontraktů (tabulka 2.3). 37 kontraktů tedy nakoupí za cenu 931, jelikož za tuto cenu se více kontraktů neprodává. Zbývajících 18 kontraktů nakoupí za cenu 932. Na ceně 932 se tedy sníží počet prodávaných kontraktů z 29 na 11. Za cenu 931 nyní nikdo žádné kontrakty nenabízí. Cena *ask* se tedy zvýšila na cenu 932 stejně jako cena *last*. Spread je nyní 2. Na poslední tabulce 2.4 lze vidět, že do trhu přišly limitní příkazy na nákup 15 kontraktů za cenu 931. Cena *bid* se tedy zvýšila na 931.

2.2 Grafy

Základem pro obchodování jsou samozřejmě data. Primárně obchodníka zajímá cenový vývoj sledovaného produktu v čase. Tento vývoj se běžně zobrazuje v grafech. Jelikož se cena produktu v čase velmi rychle mění¹, obchodníci používají grafy, které zobrazují data agre-

¹Toto je důsledek velkého množství prováděných obchodů.

Typ	Počet	Cena
Ask	2	934
	8	933
	29	932
	37	931
Bid	44	930
	25	929
	12	928
	3	927

Tabulka 2.1: Výchozí stav

Typ	Počet	Cena
Ask	2	934
	8	933
	29	932
	37	931
Bid	32	930
	25	929
	12	928
	3	927

Tabulka 2.2: Obchodník prodal 12 kontraktů

Typ	Počet	Cena
Ask	2	934
	8	933
	11	932
	32	930
Bid	25	929
	12	928
	3	927

Tabulka 2.3: Obchodník nakoupil 55 kontraktů

Typ	Počet	Cena
Ask	2	934
	8	933
	11	932
	15	931
Bid	32	930
	25	929
	12	928
	3	927

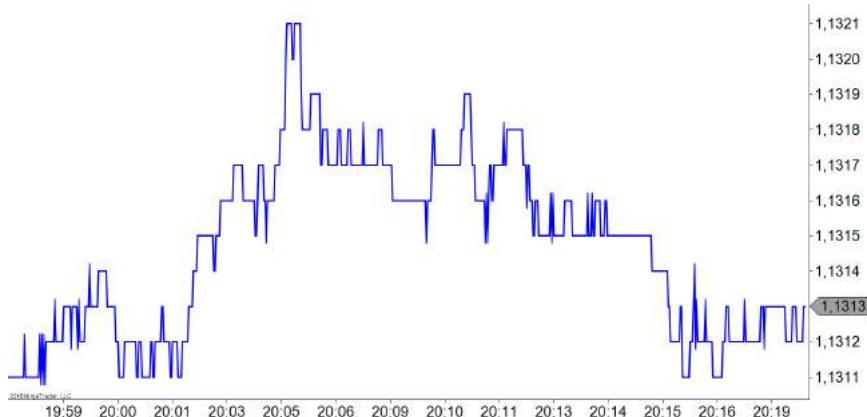
Tabulka 2.4: Přibyly limitní příkazy na nákup 15 kontraktů za cenu 931

govaná. Na obrázku 2.1 je graf vývoje ceny měnového páru EURUSD z 10.6.2015. Jedná se o tickový graf². Pokud by tento graf zobrazoval delší časové období, byl by velmi nepřehledný. Proto se data agregují. Agregovaný graf může být například denní. Údaj pro jeden den je pak určena čtyřmi cenami – *open*, *high*, *low* a *close*. Cena *open* je cena produktu na začátku dne, cena *high* představuje nejvyšší dosaženou cenu během dne, cena *low* značí nejnižší dosaženou cenu během dne a cena *close* je cena produktu na konci dne. Pro přehledné vykreslení všech 4 hodnot v každém dni je velmi často používán oblíbený svíčkový graf. Příklad takového grafu je zobrazen na obrázku 2.2.

Knot svíčky značí cenu *high* a *low*. Tělo svíčky je zase určeno cenami *open* a *close*. Pokud je cena *close* vyšší než cena *open*, pak se tělo svíčky vybarví zeleně (vzrůst ceny). V opačném případě je tělo červené (pokles ceny). Toto celé je ilustrováno na obrázku 2.3.

Kromě běžného časového grafu existuje samozřejmě více typů grafů. Jsou to grafy založené na objemu obchodů, na výchylce ceny apod. Způsoby vykreslení se také velmi různí. To implikuje spoustu možností, jak zobrazit cenový vývoj produktu. Pro účely této práce si však zcela vystačíme se základními časovými grafy.

²Tick reprezentuje jeden provedený obchod (který se ale může skládat z více kontraktů). Tickový graf tedy zobrazuje cenu produktu při každém obchodu v závislosti na čase. Z toho plyně, že hodnoty jsou v čase distribuovány nerovnoměrně.



Obrázek 2.1: Tickový graf – EURUSD, 10. 6. 2015



Obrázek 2.2: Svíčkový minutový graf – EURUSD, 10. 6. 2015

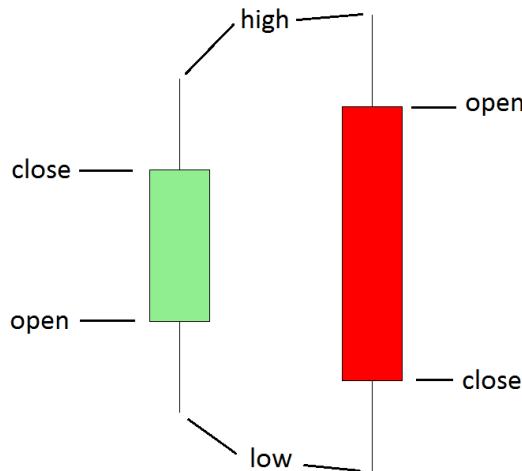
2.3 Typy produktů

V nynější době burzovní i mimoburzovní trhy nabízí nespočet produktů, se kterými lze obchodovat. Všechny produkty mají svá specifika, vyžadují různě velký kapitál, pro jejich obchodování existují zvláštní pravidla apod. V této kapitole budou rozebrány a popsány nejzajímavější produkty.

2.3.1 Forex

Pod pojmem Forex (**F**oreign **E**xchange), zkráceně FX, označujeme mimoburzovní trh měn. Produkty jsou nabízeny ve formě měnových páru. Jedná se tedy vždy o kupu a prodej určité měny. Forex je nyní nejlikvidnější trh světa. Každý den se na něm provedou obchody v celkové hodnotě až 5 bilionů dolarů [3]. Obchoduji se na něm 24 hodin denně, 5 dní v týdnu.

Obchodník obchodusící na Forexu se tedy snaží vydělat na pohybu měn. Jelikož síla měny se mění relativně pomalu, nedochází k příliš velkým výkyvům. Proto by obchodník musel nakoupit velký objem měny, aby byl jeho zisk zajímavý. To v dnešní době není



Obrázek 2.3: Popis svíček

potřeba. Forexoví brokeri nabízejí svým klientům využití pákového efektu. V tomto případě složí klient u brokera pouze zálohu, tzv. *margin*, a za to je mu umožněno disponovat s větší částkou než jakou má na účtu. Díky obchodování s vyšší částkou lze tedy dostatečně vydělávat i na menších pohybech.

Jelikož brokeři dovolují svým klientům obchodovat i s nižšími objemy měny, je tento trh finančně dostupný téměř každému. K obchodování na Forexu se lze dostat už i s 500 Kč. Avšak šance na úspěch je s touto částkou velmi malá. Výhoda tkví hlavně v tom, že obchodník může spekulovat v různém časovém horizontu. Může držet pozici (otevřený obchod) v řádu minut, ale klidně i v řádu let, a to nezávisle na jeho kapitálu.

Forexový trh je, jak už bylo řečeno, trh mimoburzovní, a tedy není centrálně regulovaný. Trh je tedy síť brokerů, tvůrců trhu, bank, fondů, pojišťoven, individuálních obchodníků apod. Tento fakt má důležité implikace. Mezi hlavní, pro individuální obchodníky, patří to, že jejich broker je jim často v obchodu protistranou. Jelikož je to jejich broker, tzv. jim „vidí do karet“. Vidí, kde mají nastavení stop lossy³, profit targety⁴ apod. Výběr důvěryhodného brokera je proto pro obchodníka klíčový.

2.3.2 Indexy

Indexy slouží jako statistický ukazatel síly určitého trhu. Mezi nejzajímavější patří akciové indexy. Tyto indexy měří sílu akciového trhu. V akciovém indexu mohou být zahrnutы akcie všech firem na trhu nebo pouze vybraných firem (například 100 největších). Jejich podíl na vývoji indexu je dán bud:

- Cenou akcií – v tomto případě se jedná o cenově vážené indexy. Podíl firmy je určen podle ceny jejích akcií.
- Tržní kapitalizací – při tomto způsobu se bere v úvahu jak cena akcií, tak i jejich

³Stop loss je maximální definovaná ztráta, kterou si určuje obchodník. Jakmile obchod této ztráty dosáhne, obchod je ukončen a obchodník tedy utržil ztrátu.

⁴Profit targetem označujeme zisk, na který obchodník míří. Při dosažení tohoto zisku je obchod ukončen a obchodník si připisuje zisk

počet. Tím je také tržní kapitalizace dána. Tento způsob lépe reflektuje velikost firem a proto se i častěji používá.

Investorům a fondům slouží akciové indexy často jako měřítko úspěšnosti nebo spíše je to nástroj pro porovnání jejich výkonnosti. Obecně překonání výnosu akciového indexu je považováno za dobrý výsledek.

Mezi nejznámější akciové indexy patří:

- Standard & Poor's 500 (S&P 500) – tento index je souborem akcií 500 vybraných, převážně amerických, firem. Je proto často používán jako ukazatel síly amerického akciového trhu. Podíl firem na tvorbě indexu je dán jejich tržní kapitalizací [6].
- Dow Jones Industrial Average – tento americký index zahrnuje 30 největších firem především z průmyslového odvětví obchodovaných na burzách NYSE a NASDAQ. Jeho počátky sahají až do roku 1896, kdy se datuje jeho vznik [9]. Je to jeden z mála indexů, který počítá podíl firem pouze na základě ceny jejich akcií.
- Nasdaq Composite – tento index je tvořen všemi firmami, jejichž akcie jsou obchodovány na burze NASDAQ. Aktuálně je to asi přes 3000 firem. Podíl firem je dán jejich tržní kapitalizací.
- Nikkei 225 – jak už název napovídá, index Nikkei se skládá z akcií 225 firem. Index Nikkei je japonský index, takže se jedná o japonské firmy obchodované na japonské burze Tokyo Stock Exchange.

Drobný investor (spekulant) provádí na indexech většinou krátkodobé až střednědobé obchody. Tyto trhy oplývají velikou likviditou, a proto jsou na krátkodobé obchodování více než vhodné. Obchodníci, podobně jak u Forexu, mohou využít pákového efektu, což dělá indexy populárními obchodními produkty mezi menšími obchodníky.

2.3.3 Akcie

Akcie neboli cenný papír představuje určitou část společnosti (firmy). Vlastník akcie tedy vlastní část společnosti a bývá označován za akcionáře firmy. Jako akcionář firmy má právo na podíl z hospodářského zisku společnosti, tzv. dividendu. Pokud firma zisku nedosáhne, dividendu samozřejmě nevyplácí. Dividenda bývá obvykle vyplácena (pokud vůbec) každý rok a o její výši rozhoduje valná hromada. Dividendový výnos, který akcionář získá, je přímo úměrný počtu akcií, které vlastní. Kromě tohoto výnosu může akcionář dosáhnout ještě kapitálového výnosu. Tento výnos plyne z prostého prodání akcie/akcií [33]. V České republice může být akcionář osvobozen od daně z kapitálového výnosu, pokud splní tzv. časový test. Časový test v překladu znamená to, že pokud je akcie držena méně než 3 roky, je obchod považován za spekulaci s cílem zisku, a proto se v tomto případě platí daň z příjmu. Pokud je akcie držena déle, při prodeji je zisk z prodeje od daně osvobozen. Doba 3 roky je platná od 1. 1. 2014. Předtím byla tato doba dlouhá pouze 6 měsíců. O časovém testu pojednává §4, zákon č. 586/1992 Sb. o daních z příjmů [35]. U dividend je obvyklá daň 15%.

Z hlediska spekulací je vhodné držet akcie v řádu měsíců, ale spíše let. Jedná se tedy o dlouhodobé investice. U akcií nelze většinou využít pákového efektu. Někteří brokeři sice pákový efekt nabízí, ale ten je většinou malý ve srovnání například s indexy. Proto je potřeba mít na obchodování akcií dostatečně velký kapitál (ideálně miliony).

Poslední specifikum akcií je směr spekulace. U obvyklých produktů je běžné spekulovat na růst i pokles. U akcií je běžná spekulace pouze na růst. Spekulace na pokles je možná pouze u některých obchodovaných firem, a to v závislosti na brokerovi. Pro spekulaci na pokles je totiž potřeba si akcii „půjčit“. Většinou to proto probíhá tak, že si obchodník akcii půjčuje od svého brokera. Ten musí mít tedy dané akcie koupené. Navíc si broker při půjčení akcií účtuje určitý úrok.

2.3.4 Komodity

Mezi komodity se řadí kovy (zlato, měď, platina, …), zemědělské produkty (sójové boby, kukuřice, pšenice, …), maso (vepřové maso, hovězí maso, …) a energie (ropa, zemní plyn, benzín, …). Neobchoduji se přímo s fyzickými produkty, ale pouze s kontrakty na tyto produkty. Kontrakt je vlastně smlouva o tom, že obchodník danou komoditu k datu vypršení konaktu zakoupí za cenu stanovenou v době zakoupení konaktu. Proto je potřeba se zakoupenému konaktu zbavit před datem vypršení konaktu. Pro každou komoditu je její kontrakt pevně definován. Například kontrakt bílého cukru se vztahuje na 1 tunu cukru. Nejmenší cenový pohyb je potom \$5 [17].

U komodit lze opět využít pákového efektu.

2.3.5 Opce

Koupí opce si obchodník zajistí právo (ne povinnost) na nákup nebo prodej specifikovaného produktu za předem stanovenou cenu. Využít toho lze pouze do té doby, než opce vyprší (termín expirace). Cena opcí (označována jako *opcni prémium*) se velmi liší v závislosti na termínu expirace opce a vzdálenosti stanovené ceny od aktuální ceny produktu na trhu. V konečném důsledku běžní obchodníci využívají opce často jako obchodovaný produkt. Snaží se tedy opci levně kupit a draze prodat.

Obchodování opcí vyžaduje poměrně vyšší kapitál a nabízí nižší zhodnocení než ostatní produkty. Obchodování opcí se také principiálně velmi liší od klasického obchodování, a proto se obchodováním opcí v této práci věnovat nebudeme.

2.3.6 Spready

Při obchodování spreadů se obecně pracuje se dvěma produkty. Spekuluje se totiž na rozdíl ceny mezi produkty. Pro obchodníka tedy není důležité, zda trhy rostou nebo klesají, zajímá ho jen rozdíl mezi konkrétními komoditami.

Spready jsou obecně považovány za bezpečnou investici, která je ale vykoupena nižším zhodnocením. Jelikož je tento typ obchodování opět rozdílný od toho klasického, nebude mu v této práci věnována pozornost.

2.4 Analýzy

V této kapitole budou rozebrány základní typy obchodník analýz, které byly naznačeny již v úvodu.

2.4.1 Technická analýza

Technická analýza spočívá v analýze historického vývoje ceny produktu. V této analýze nemusí být zahrnuta pouze cena, ale třeba i objem obchodů (*volume*). Způsobů, jak tyto

veličiny zkoumat a podle nich určovat budoucí vývoj ceny, je nespočet. Zcela základním nástrojem jsou klouzavé průměry a další běžné indikátory⁵. Obchodníci kromě toho také často kreslí do grafu. Vyznačují si *supporty* a *resistence*⁶, kreslí trendové linky, významné cenové úrovně atd. Mezi pokročilejší techniky pak patří například intermarket analýza, kdy obchodník porovnává vývoj ceny s vývojem ceny jiného produktu a hledá mezi těmito produkty určité závislosti. Vzhledem k tomu, že k provedení technické analýzy stačí obchodníkovi pouze graf, těší se technická analýza velké popularitě.

2.4.2 Fundamentální analýza

Fundamentální analýza se snaží podchytit vliv událostí reálného světa na vývoj ceny sledovaného produktu. Předpokládá, že vývoj ceny tyto události významně ovlivňuje. Například lze očekávat, že díky velkému suchu v zemích, kde se pěstuje kukuřice, poroste kvůli neúrodě cena kukuřice nahoru. Obchodníci provádějící fundamentální analýzu se tedy snaží analyzovat předem vytipované zprávy a podle nich se rozhodují, zda budou kupovat, prodávat nebo ani jedno z toho. Kupříkladu sledují finanční reporty firem, zprávy ze zasedání Fedu⁷, míru úrokové sazby a další ekonomické zprávy a faktory. Vzhledem k tomu, že analýza všech těchto zdrojů a informací je časově velmi náročná, vyžaduje neustálý přehled o ekonomické situaci, orientaci v oboru a zkušenosti, je tento druh analýzy obecně vnímán jako pokročilejší. Dá se říci, že její význam roste s časovým rámcem, ve kterém obchodník obchoduje.

Tato práce se bude věnovat právě fundamentální analýze. Budou zpracovávány zdroje informací pouze **kvantitativního charakteru**. Informace čerpaná z „čísel“ je totiž jednoznačně interpretovatelná a její zpracování tedy významně jednodušší. Technická analýza bude do tohoto procesu použita pouze v tom smyslu, že budeme pracovat s historickým vývojem ceny, což je nevyhnutelné pro testování obchodního systému.

Ovlivňující faktory

V této podkapitole budou uvedeny zdroje informací, ze kterých může být čerpáno a které mohou mít vliv na vývoj ceny určitých produktů.

- Historický vývoj ceny produktu — Tento zdroj informací je sice nástrojem technické analýzy, ale jeho použití je nutné k analýze a testování obchodního systému. Proto se s historickým vývojem ceny běžně pracuje i ve fundamentální analýze.
- Počasí — Počasí může ovlivňovat velmi významně cenu hlavně komodit z kategorie zemědělských produktů – kukuřice, pšenice, ječmen apod. Kromě komodit může ovlivňovat také náladu a rozpoložení lidí, kteří na burze obchodusí.
- Války a přírodní katastrofy — Z vysokého počtu obětí (nestandardních úmrtí) lze odvodit například válku nebo přírodní katastrofu. Tyto události mohou mít velký vliv například na akcie firem sídlících v takto postižených oblastech. Naopak může zvýšit cenu firem, které z těchto událostí těží (zbrojní průmysl).
- Ceny souvisejících komodit — Vysoká cena některých komodit může velmi ovlivňovat firmy, které na těchto komoditách stojí. Například zvýšena cena kakaových bobů může stát za zvětšením výdajů firmy Nestlé.

⁵RSI, MACD, Williams %R, CCI, Bollinger Bands apod. [44][42][45]

⁶Místa s velkým střetem nabídky a poptávky.

⁷Jako Fed (federální rezervní systém) je označován centrální bankovní systém USA [4].

- Volební preference — Strana, která je populární a má většinu při rozhodování o zájomech, má také moc ovlivňovat ceny produktů. Pokud se bude jednat o ekologickou stranu, jak to ovlivní ceny benzínu? Pokud strana bude militaristicky založená, jak to ovlivní zbrojařské firmy?
- Počty uprchlíků — Počet uprchlíku (například v USA) může být příčinou nebo zdrojem velkých změn. Tyto změny můžou mít velký vliv na ekonomiku.
- Rozpočty, dotace — Z výše plánovaných rozpočtů lze posoudit určitou ekonomickou situaci. A pokud dostane firma dotace, pravděpodobně je využije ke svému růstu. Naopak pokud firma potřebuje dotace, aby se *udržela nad vodou*, není zřejmě dostatečně silná. Důležité je, že dotace mohou mít vliv na hodnotu firmy.
- Hodnota konkurenčních firem — Hodnota firmy (akcií) může korelovat s hodnotou firem podnikajících ve stejném segmentu. Z poklesu akcií firmy lze například usoudit, že se spíše prosazuje konkurence. Případně pokles akcií více firem ve stejném segmentu může indikovat celkový pokles segmentu, a tedy lze očekávat pokles akcií i ostatních firem.
- Počet obětí výrobku — Jak ovlivní ceny akcií firmy Glock počet obětí jejich výrobků? Nebo souvisí počet obětí/autonehod aut od určité značky s cenou akcií této značky?
- Počet zaměstnanců — Ze stálého nabírání nových zaměstnanců můžeme soudit, že firmě se daří a má dobré vyhlídky. Naopak masové propouštění je pravděpodobně následkem problémů firmy.
- Finanční reporty firem — Z finančních reportů firem lze odvozovat jejich ekonomický stav. Sledovat lze hlavně výsledky hospodaření, výše vyplácené dividendy investorům apod.
- Kurz měn — Změna kurzu měny, která je primární pro firmu, může ovlivňovat cenu jejích výrobků v zahraničí. A to může mít zase vliv na hodnotu firmy. Může být tedy vhodné sledovat kurzy nejobchodovanějších měn.
- Počet prodaných výrobků — Počet výrobků, které firma prodá, může být indikátorem ekonomického stavu firmy. Pokud přestává být o její výrobky zájem, firma musí přijít s něčím novým nebo její zisky budou stále klesat a firma bude postupně upadat.
- Hospodaření států — Z hospodaření státu (výše rozpočtu, dluh) lze odhadnout ekonomickou stabilitu státu. Jak toto může ovlivnit firmy, které operují hlavně na trhu tohoto státu? Například při zvyšujícím se dluhu státu, může mít stát snahu vybírat více peněz na daních a tím více zatěžovat firmy v tomto státě podnikající.
- Roční období — Roční období souvisí hodně s počasím, přesto bude rozepsáno zvlášť. Roční období bude určitě ovlivňovat minimálně cenu zemědělských komodit.
- Cena akcií dodavatelských firem — Firmy stojí na svých dodavatelích, obzvláště když mezi dodavateli není mnoho alternativ. Takže lze očekávat vliv hodnoty firmy Intel na akcie firmy Lenovo, pokud majoritu počítačů firmy Lenovo pohání právě procesory Intel.

- Počet výsledků v Google vyhledávání — Při vyhledání názvu firmy v Google lze z počtu nalezených výsledků odvozovat „popularitu firmy“. Dle výše zmíněné práce [36] o souvislosti Google Trends a hodnoty akcií by zde měla existovat závislost.
- Počet zobrazení nebo editací článku na Wikipedii — Souvislost je zde podobná jako u Google vyhledávání. A opět by zde dle výše zmíněné práce [25] tato souvislost měla existovat.

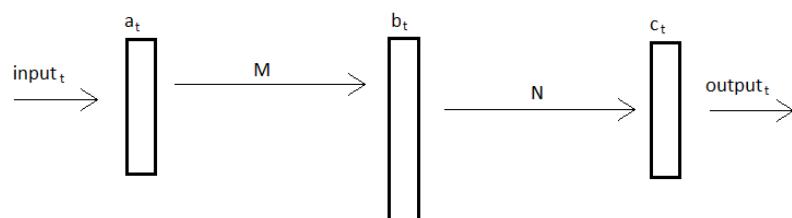
Kapitola 3

Neuronové sítě

K predikci vývoje cen bude použito strojové učení, konkrétně neuronová síť. Neuronová síť se běžně (a s úspěchem) používá na rozpoznávací úlohy, predikci hodnot, generování dat apod. Jako příklady úspěšného použití lze zmínit – neuronová síť jako jazykový model [23], neuronová síť jako prediktor krachu banky [2] nebo neuronová síť jako rozpoznávač poznávacích značek na autech [50].

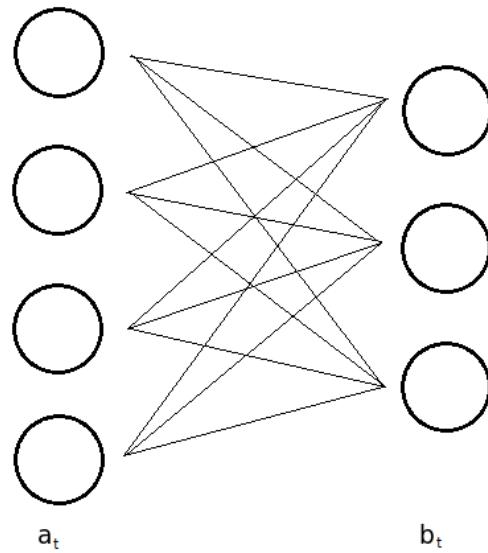
3.1 Základní model neuronové sítě

Neuronová síť je matematický model s inspirací v lidském mozku. Jeho základem jsou neurony, které jsou uspořádány do vrstev. Toto schéma můžeme vidět na obrázku 3.1. Základní architektura je tvořena vstupní vrstvou, skrytou vrstvou a výstupní vrstvou. V každé vrstvě je předem daný a pevně stanovený počet neuronů. V sousedních vrstvách je každý neuron spojen s každým neuronem ze sousední vrstvy (obr. 3.2). Mezi neurony jsou tedy spoje, přitom každý spoj má nějakou váhu. Jako *blackbox* funguje neuronová síť tak, že do vstupní vrstvy je nahrán vstup. Následně se spustí výpočet, který počítá od vstupní vrstvy postupně až k výstupní. Z výstupní vrstvy pak lze zkopirovat vypočítaný výsledek. Podrobněji bude tento výpočet popsán v následující podkapitole.

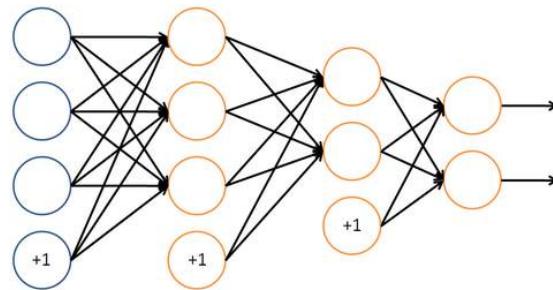


Obrázek 3.1: Základní architektura neuronové sítě [16]

Tato základní architektura bývá často rozšířena o bias neurony. Takto upravená architektura je zobrazena na obrázku 3.3. Každá vrstva (kromě té výstupní) je rozšířena o jeden neuron. Tento neuron není připojen k žádné předchozí vrstvě, ale s následující vrstvou je spojen stejně jako ostatní neurony v dané vrstvě. Zvláštností je, že v průběhu celého výpočtu má konstantní hodnotu. Jeho váhy se ale během trénování mění stejným způsobem jako u ostatních neuronů. Bias neurony zlepšují trénování modelu tak, že jsou schopny hodnoty neuronů v následující vrstvě posunout do jiného (vhodnějšího) intervalu hodnot.



Obrázek 3.2: Propojení neuronů mezi vrstvami [16]



Obrázek 3.3: Propojení neuronů mezi vrstvami s bias neurony [19]

3.1.1 Princip výpočtu

Popis výpočtu je částečně převzat z [16] a [23]. Pracujme s architekturou neuronové sítě, jaká je na obrázku 3.1. Pro vysvětlení výpočtu definujme výpočet chyby a dva základní algoritmy – dopředný průchod a zpětný průchod. Jejich princip se uplatňuje v každé neuronové síti.

Dopředný průchod

Dopředný průchod, známý spíše jako *forward pass*, definuje výpočet výstupní vrstvy z vrstvy vstupní. Pracuje následovně.

Do vstupní vrstvy se v čase t nahraje vybraný vstupní vektor \mathbf{a} o velikosti $|\mathbf{a}|$. Následně se pomocí hodnoty neuronu ve vstupní vrstvě a váhy spoje mezi odpovídajícími neuronami spočítají hodnoty neuronů ve skryté vrstvě \mathbf{b}_t :

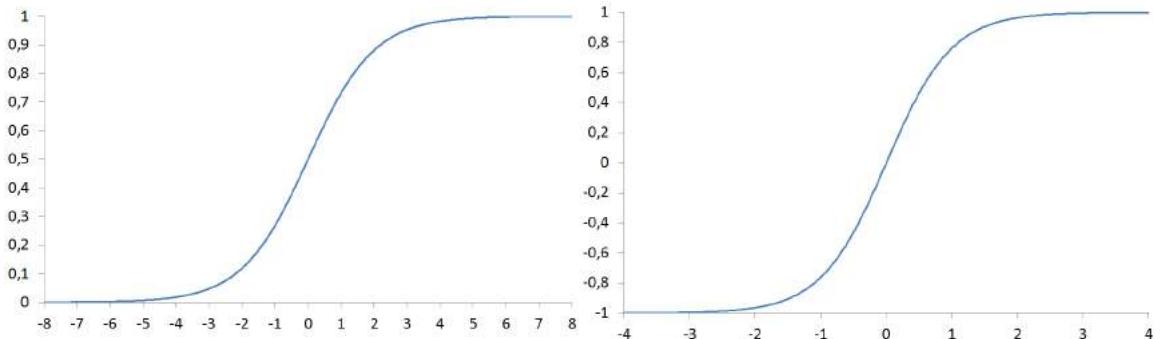
$$\mathbf{b}_t^j = f \left(\sum_i a_t^i M^{ji} \right) \quad (3.1)$$

Přitom funkce $f(x)$ může být například sigmoid (3.2, obr. 3.4) nebo tanh (3.3, obr. 3.5):

$$f(x) = \frac{1}{1 + e^{-x}} \quad (3.2)$$

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (3.3)$$

Dle [20] je vhodné použít symetrickou funkci se střední hodnotou 0, a proto lze zvolit například tanh.



Obrázek 3.4: Sigmoida

Obrázek 3.5: Tanh

Následně se obdobným způsobem jak v 3.1 spočítají hodnoty neuronů ve výstupní vrstvě:

$$\mathbf{c}_t^k = g \left(\sum_j \mathbf{b}_t^j \mathbf{N}^{kj} \right) \quad (3.4)$$

Zde funkce $g(x)$ může být například softmax (3.5), který zajistí, že součet hodnot neuronů ve výstupní vrstvě bude 1. To se využije zpravidla pokud ve výstupní vrstvě mají být pravděpodobnosti.

$$g(z_i) = \frac{e^{z_i}}{\sum_k e^{z_k}} \quad (3.5)$$

Po tomto kroku je výsledek (odezva) neuronové sítě na zadáný vstupní vektor uložen ve výstupní vrstvě \mathbf{c}_t .

Výpočet chyby

Výpočet chyby se provádí porovnáním dvou vektorů. Konkrétně se porovnává vypočítaný vektor ve výstupní vrstvě a očekávaný vektor:

$$\mathbf{e}_t^o = \mathbf{d}_t - \mathbf{c}_t, \quad (3.6)$$

kde \mathbf{d}_t je hodnota požadovaného výstupu. Výše zmíněný výpočet chyby se použije v následujícím kroku – *backpropagation*.

Kromě toho se je potřeba zmínit hodnotící funkci. To je funkce, jejíž hodnota představuje *přesnost* neuronové sítě. Při učení neuronové sítě se snažíme právě tuto hodnotu minimalizovat. Oblíbenou hodnotící funkcí je MSE¹, který se spočítá následovně:

$$mse = \frac{1}{L} \sum_{i=1}^L (\mathbf{d}_t^i - \mathbf{c}_t^i)^2, \quad (3.7)$$

kde L je velikost výstupního vektoru.

Zpětný průchod

Zpětný průchod, známý též jako *backpropagation*, definuje adaptaci vah podle chyby uložené ve výstupní vrstvě. Váhy se upravují od výstupní vrstvy směrem k vrstvě vstupní. Cílem úpravy vah je minimalizace chyby.

Nejprve se upraví váhy spojů mezi skrytou a výstupní vrstvou:

$$\mathbf{N}_{t+1}^{jk} = \mathbf{N}_t^{jk} + \mathbf{b}_t^j \mathbf{e}_t^{ok} \alpha - \mathbf{N}_t^{jk} \beta \quad (3.8)$$

kde α je *learning rate* a β je parametr regularizace (viz [23]). Podle nových vah se spočítá chyba ve skryté vrstvě:

$$\mathbf{e}_t^h = d^h((\mathbf{e}_t^o)^T \mathbf{N}, t) \quad (3.9)$$

$$d^h(\mathbf{x}, t) = \mathbf{x} \mathbf{b}_t^j (1 - \mathbf{b}_t^j) \quad (3.10)$$

Zbývá upravit váhy mezi vstupní a skrytou vrstvou:

$$\mathbf{M}_{t+1}^{ij} = \mathbf{M}_t^{ij} + \mathbf{a}_t^i \mathbf{e}_t^{hj} \alpha - \mathbf{M}_t^{ij} \beta \quad (3.11)$$

¹Mean squared error

Kompletní algoritmus

První fázi je učení neuronové sítě. V této fázi jsou na vstup (vstupní vrstva) neuronové sítě postupně nahrávány trénovací vstupní vektory. Pro každý trénovací vektor se provede nejdříve *dopředný průchod*. Výsledek neuronové sítě ve výstupní vrstvě je poté porovnán s očekávaným výstupem velikosti $|c|$, který je specifikován pro zadaný vstupní vektor. Vy počte se tedy chyba požadovaného výstupu oproti vektoru ve výstupní vrstvě. Tato chyba se uloží do výstupního vektoru. Následuje *zpětný průchod*, při kterém se adaptují váhy podle zjištěné chyby. Tímto je dokončen proces zpracování jednoho vstupního vektoru. Neuronová síť takto postupně zpracovává všechny trénovací vektory. Jakmile jsou všechny trénovací vektory zpracovány, nastává validační fáze.

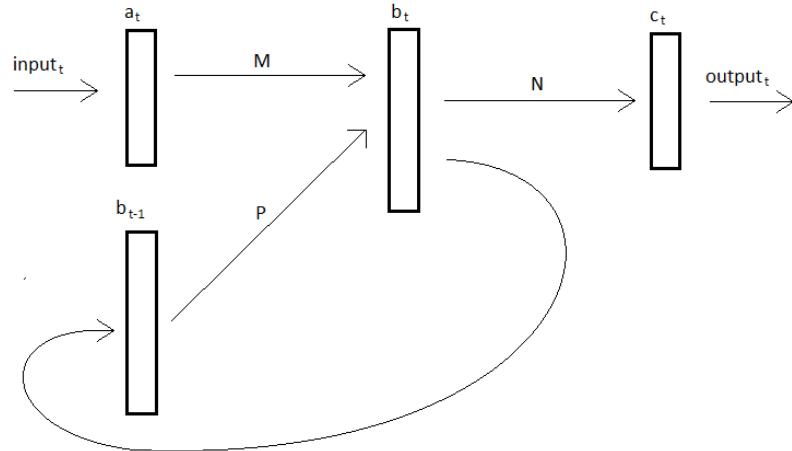
Ve validační fázi neuronová síť zpracovává validační vektory. Počet těchto vektorů bývá zhruba desetinový oproti počtu trénovacích vektorů. Při zpracování validačních vektorů je prováděn pouze *dopředný průchod* a výpočet chyby, tedy neuronová síť se ze vstupů neučí. Tato fáze se používá ke spočítání celkové chyby neuronové sítě na neviděných datech. Pokud je tato chyba stále „velká“, dochází k další iteraci (trénování a validace). Rozhodnutí, že už není chyba příliš „velká“ a tedy je trénování ukončeno, může být v praxi prováděno různě. Lze použít například pevný počet iterací, ukončit trénování pokud rozdíl chyby z aktuální iterace a předchozí iterace je menší než nastavený práh apod. Splnění této podmínky nemusí vést nutně k okamžitému zastavení trénování. Lze ještě zmenšovat parametr α a trénování dále zpřesňovat. Po ukončení trénování je vhodné model uložit pro vícenásobné použití.

Nakonec dochází k testování sítě nebo jejímu praktickému využití. Neuronové sítě se na vstup přiloží testovací vektor. Následně neuronová síť provede *dopředný průchod* a tím spočítá *odezvu* hodnotu výstupního vektoru. Poté lze propagovat chybu zpět sítí a síť tak při testování učit, ale pouze pokud je správná hodnota výstupního vektoru známá. Takto se síť postupně předloží všechny testovací vektory.

3.2 Typy neuronových sítí

Architektura běžně používaných neuronových sítí se různí. Můžou se lišit počtem skrytých vrstev, zapojením, uspořádáním apod. V této podkapitole budou popsány některé často používané architektury

- Dopředná neuronová síť (feed forward neural network) — Jedná se o základní architekturu neuronové sítě. Na této architektuře byl popisován algoritmus učení, validace a testování neuronové sítě v podkapitole 3.1.1.
- Rekurentní neuronová síť (recurrent neural network) — Tento speciální typ architektury má ve vstupní vrstvě kopii skryté vrstvy z času $t - 1$ (obr. 3.6). Její předností je zachování kontextu vstupních vektorů. Zatímco dopředná neuronová síť na stejný vektor odpoví vždy stejným výstupem, rekurentní neuronová síť odpoví různými výstupy v závislosti na kontextu (předchozích vstupních vektorů). Při predikci je tedy výhodné použít tuto architekturu, protože není třeba definovat pevný počet předchozích vzorků signálu. Podstatou výpočtu rekurentní neuronové sítě je prakticky stejná způsobem jak u dopředné neuronové sítě (s ohledem na architekturu).
- Hluboká neuronová síť (deep neural network) — Tato neuronová síť je specifická tím, že obsahuje dvě a více skrytých vrstev neuronů. Jejich trénování už vyžaduje pokročilejší techniky a přístupy (genetické algoritmy, *pre-training*) [11]. Teprve až



Obrázek 3.6: Architektura rekurentní neuronové sítě [16]

po použití těchto technik se hluboké neuronové sítě staly zajímavými [14]. Jejich nevýhodou je velká výpočetní náročnost, proto je snaha výpočet přesunout co nejvíce na GPU.

- Konvoluční neuronová síť (convolutional neural network) — Konvoluční neuronová síť se skládá z několika konvolučních vrstev následovaných několika plně propojenými vrstvami jako v klasických neuronových sítích. Tato architektura je vhodná na zpracování 2D dat (obraz) [7]. Vzhledem k tomu, že se v této práci nebude s takovými daty pracovat, není třeba se touto architekturou dále zabývat.

3.2.1 Použití

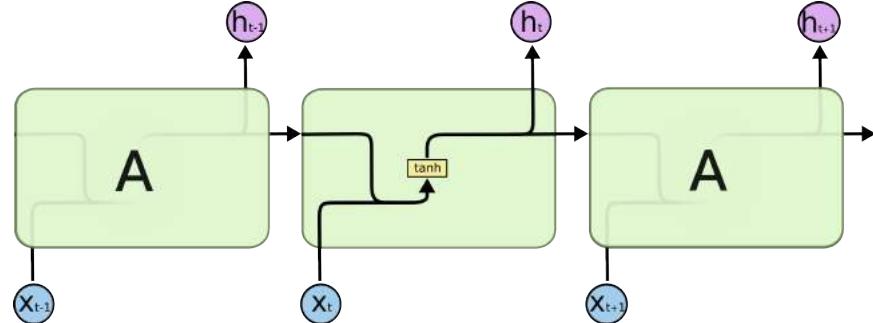
Data neuronové sítě jsou sekvence dvojic vektorů – vstupního a očekávaného výstupního. Neuronová síť se tedy učí správné reakce na vstupní vektor. Mezi nejčastější aplikace proto patří klasifikace a regrese (predikce). Klasifikovat jde takto téměř cokoliv, je nutné pouze data převést do formátu vhodného pro neuronovou síť (m -dimenzionální vektor). V této práci bude síť používána jako prediktor. Cílem tedy bude predikce budoucích hodnot libovolného signálu – zde bude signálem vývoj ceny v čase.

3.3 LSTM neuronová síť

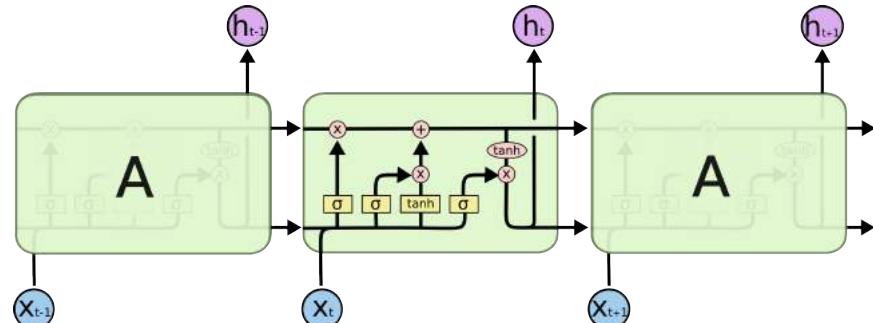
LSTM (long-short term memory) neuronová síť je rekurentní neuronová síť se specifickou architekturou. Výše zmíněná běžná rekurentní neuronová síť se učí dobře závislosti s krátkým „časovým“ rozestupem. Například závislost mezi vektory v časech t a $t - 1$ se naučí lépe než mezi vektory v časech t a $t - 200$. Na některé úlohy je tedy běžná RNN vhodná, avšak obecně je LSTM neuronová síť výkonnější než běžná RNN. Rozdíl tkví právě v tom, že LSTM neuronová síť je schopná se naučit závislosti s krátkým i delším časovým rozestupem. Následující popis fungování LSTM neuronové sítě je částečně převzat z [32].

Na obrázku 3.7 se nachází schéma běžné rekurentní neuronové sítě. Na tomto obrázku je znázorněn průchod 3 vektorů. Jednotlivé bloky představují skrytu vrstvu, která může mít různě složitou strukturu. Nemusí se tedy nutně skládat z určitého počtu neuronů. Jedná se

vlastně o bloky, které jsou centrem rekurence. Prostřední blok znázorňuje skrytou vrstvu v čase t . U RNN lze vidět, že vstupem skryté vrstvy v čase t je hodnota vektoru z času t a hodnota ze skryté vrstvy z času $t - 1$. Tento vstup je přiveden na vstup vrstvy složené ze standardních neuronů s aktivační funkcí tanh. Vstup se klasicky *složí* a na výsledek se aplikuje aktivační funkce (viz vzorec 3.1). Výsledkem je vektor, který je poté distribuován do výstupní vrstvy a do vstupu skryté vrstvy v čase $t + 1$ běžným způsobem, jaký je popsán v podkapitole 3.1.1.



Obrázek 3.7: Běžná rekurentní neuronová síť [32]

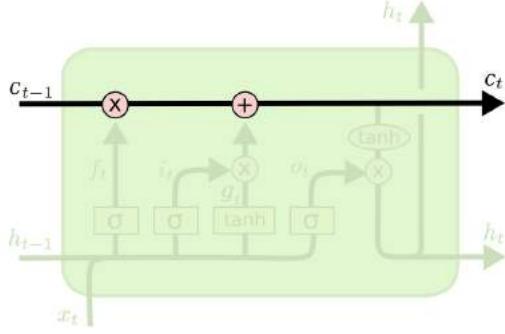


Obrázek 3.8: LSTM neuronová síť [32]

Skrytá vrstva LSTM neuronové sítě má ale strukturu složitější, jak je vidět na obrázku 3.8. Hodnota výstupního vektoru (\mathbf{h}) skryté vrstvy v čase t je na vstupu skryté vrstvy v čase $t + 1$ stejně jako v případě RNN (výpočet výstupního vektoru je ale samozřejmě jiný). Klíčem je zde vektor \mathbf{c} nazývaný jako stav buňky. Ten je uchováván ve skryté vrstvě (viz obrázek 3.9). Tento vektor prochází hradly, která ze stavu ubírají informaci (*forget gate*) nebo do něj přidávají informaci (*input gate*). Hradlo (obecně) je komponenta, jehož hodnota je mezi 0 a 1. Čím vyšší je tato hodnota, tím více informace projde.

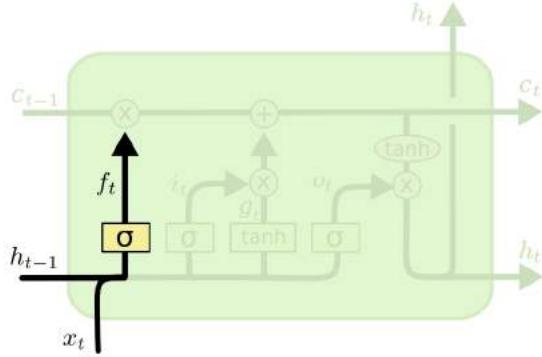
Prvním hradlem, kterým stav buňky prochází, je *forget gate*. Výpočet jeho hodnoty je znázorněn na obrázku 3.10. Výstupní vektor \mathbf{h}_{t-1} z předchozího kroku je spojen se vstupním vektorem \mathbf{x}_t . Tento vektor je přiveden na vstup vrstvy se sigmoidou jako aktivační funkcí. Výstupem je vektor \mathbf{f}_t . Tento vektor se tedy spočítá následovně:

$$\mathbf{f}_t^j = \sigma \left(\mathbf{b}_f^j + \sum_k [\mathbf{h}_{t-1}, \mathbf{x}_t]^k \mathbf{W}_f^{jk} \right) \quad (3.12)$$



Obrázek 3.9: Průchod vektoru \mathbf{c}_{t-1} neuronem [32]

Výraz $[,]$ zde značí konkatenaci vektorů, horní indexy značí vektorový index, \mathbf{b}_f je bias neuronu a \mathbf{W}_f jsou váhy vrstvy.



Obrázek 3.10: Forget gate [32]

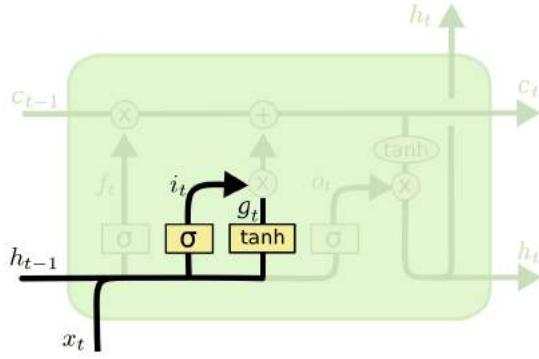
Následuje spočítání *input gate*. Toto je ilustrováno na obrázku 3.11. Opět dochází ke konkatenaci výstupního vektoru \mathbf{h}_{t-1} a vstupního vektoru \mathbf{x}_t . Konkatenovaný vektor je přiveden na vstup vrstvy se sigmoidou jako aktivační funkcí. Výsledkem bude vektor \mathbf{i}_t , který bude rozhodovat, jaké hodnoty budou přidány. Dále je konkatenovaný vektor přiveden na vstup vrstvy, který má tanh jako aktivační funkci. Výsledkem bude vektor \mathbf{g}_t obsahující hodnoty určené k přidání do stavu buňky. Vektory se tedy spočítají následovně:

$$\mathbf{i}_t^j = \sigma \left(\mathbf{b}_i^j + \sum_k [\mathbf{h}_{t-1}, \mathbf{x}_t]^k \mathbf{W}_i^{jk} \right) \quad (3.13)$$

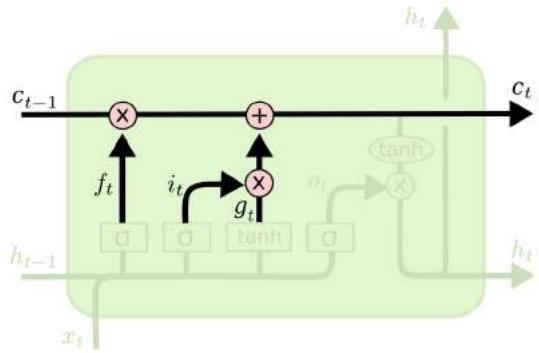
$$\mathbf{g}_t^j = \tanh \left(\mathbf{b}_g^j + \sum_k [\mathbf{h}_{t-1}, \mathbf{x}_t]^k \mathbf{W}_g^{jk} \right) \quad (3.14)$$

Nyní lze spočítat průchod vektoru \mathbf{c}_{t-1} hradly *forget gate* a *input gate*, a tedy spočítat nový stav buňky \mathbf{c}_t . Průchod je znázorněn na obrázku 3.12.

$$\mathbf{c}_t = \mathbf{c}_{t-1} \mathbf{f}_t + \mathbf{g}_t \mathbf{i}_t \quad (3.15)$$



Obrázek 3.11: Input gate [32]



Obrázek 3.12: Průchod stavu buňky hradly [32]

Posledním krokem je spočítání výstupního vektoru \mathbf{h}_t , jak je znázorněno na obrázku 3.13. Tato část se nazývá *output gate*. V tomto případě se konkatenovaný vektor přiloží na vstup vrstvy se sigmoidem. Výstupní vektor \mathbf{o}_t určí, které hodnoty ze stavu buňky půjdou na výstup. Poté se vektor stavu buňky \mathbf{c}_t proloží funkcí tanh. To zajistí, že se hodnoty tohoto vektoru dostanou do intervalu $(-1; 1)$. Nakonec tento vektor projde hradlem *output gate* a výsledný výstupní vektor je odveden na výstup skryté vrstvy.

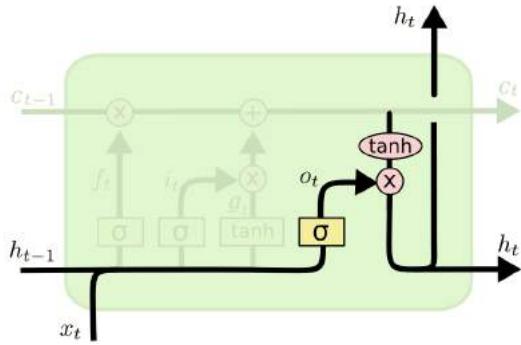
$$\mathbf{o}_t^j = \sigma \left(\mathbf{b}_{\mathbf{o}}^j + \sum_k [\mathbf{h}_{t-1}, \mathbf{x}_t]^k \mathbf{W}_{\mathbf{o}}^{jk} \right) \quad (3.16)$$

$$\mathbf{h}_t = \tanh(\mathbf{c}_t) \mathbf{o}_t \quad (3.17)$$

Učení LSTM neuronové sítě je velmi podobné učením běžných neuronových sítí [39]. Použít lze ale i pokročilejší metody [1] [26].

3.4 Dlouhodobá predikce

Vzhledem k tomu, že akcie je vhodné držet delší dobu (tzv. buy-and-hold strategie), dále kvůli snížení poplatků za obchodování a vyhnutí se tomu, že všechna používaná data nemusí



Obrázek 3.13: Output gate [32]

být dostupná ihned na konci dne, je vhodné predikovat v delším časovém horizontu než je v tomto případě 1 den. Možnosti, jak k tomuto přistoupit, je opět více.

1. První metodou je zřetězení více neuronových sítí. Tento princip je popsán v [10]. Je vytvořena kaskáda neuronových sítí, kde každá neuronová síť má kromě standardních vstupů připojeny také vstupy z predikce pro předchozí den. Z tohoto principu plyne zvýšená výpočetní náročnost až několikanásobně.
2. Další metodou je naučit neuronovou síť predikovat hodnotu pro následující den a tuto hodnotu pak v dalším kroku použít jako vstup. Tento postup pak několikrát opakovat. Výsledek této metody bude pochopitelně při delším časovém horizontu velmi nepřesný. Další nevýhodou je nutnost predikovat všechny vstupní proměnné.
3. Poslední zde uvedenou metodou je predikce nikoliv hodnoty následujícího dne, ale hodnoty například 30. dne. Jako vstup lze použít hodnoty předchozího dnu, předchozích dnů, ale i třeba hodnotu 30 dní zpět.

Všechny zmíněné metody budou dále okomentovány a některé vyzkoušeny v kapitole 8.

3.5 Dropout

Dropout je efektivní technika zabírající přeúčení neuronové sítě. Při experimentování s neuronovou sítí se přeúčení sítě dosáhne lehce a často. Stačí, pokud je síť až příliš komplexní na řešený problém, nebo trénování probíhá na nedostatečném množství dat². V takových případech se neuronová síť naučí konkrétní trénovací vektory a nikoliv rozeznávat vzory v datech, tedy generalizovat řešený problém. Jednou možností je zmenšit velikost neuronové sítě. Tento způsob ale nemusí vést k nejlepším možným výsledkům. Vhodnější je aplikovat dropout.

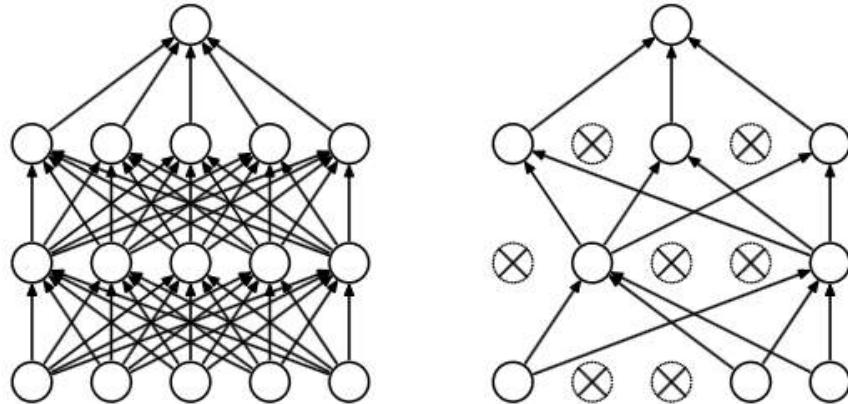
Tato technika funguje následovně. V trénovací fázi dochází k vypínání náhodně vybraných neuronů spolu s jejich spoji. Toto je ilustrováno na obrázku 3.14. Aktivní a neaktivní neurony se volí pro každý trénovací vektor zvlášť (nebo více vektorů v případě dávkové backpropagation). Pravděpodobnost, se kterou neuron zůstane aktivní, může být nalezena empiricky, ale ve většině případů je vhodné použít konstantu 0,5. V případě vstupní vrstvy bývá tato hodnota cca 0,8 [41]. Během testování už k dropoutu nedochází. Místo toho jsou

²Což je vlastně jedno a to samé.

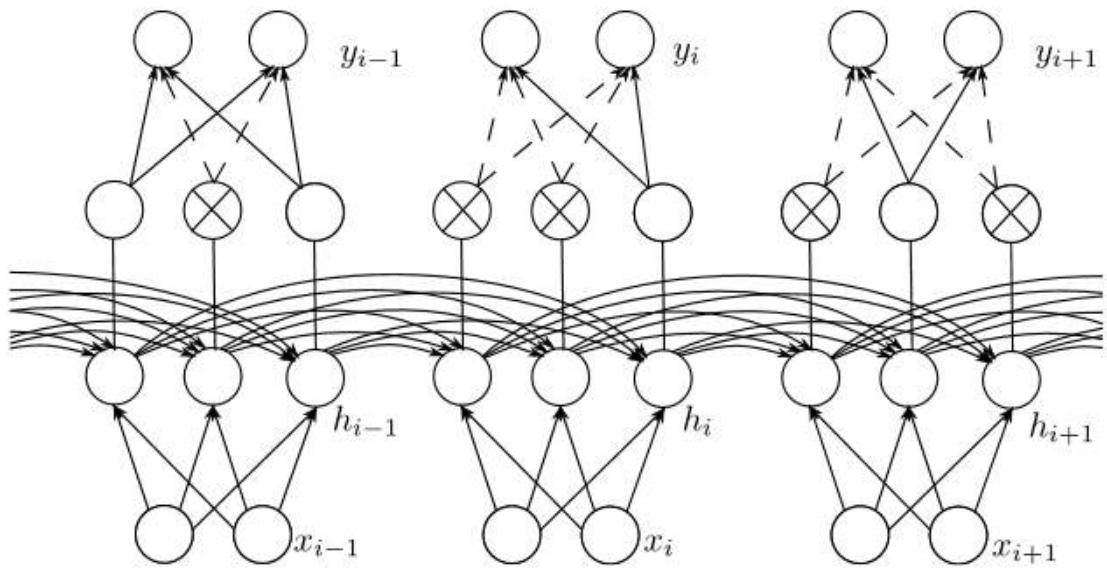
všechny váhy adekvátně zmenšeny, aby výstupní hodnoty byly ve stejných mezích jako při trénování. Při použití dropout o hodnotě 0,5 jsou tedy všechny váhy v této vrstvě zmenšeny na polovinu.

Důsledkem vypínání náhodných neuronů je skutečnost, že s každým trénovacím vektorem dochází k učení jiné neuronové sítě. Pokud je neuronová síť složena z n neuronů, existuje až 2^n různých neuronových sítí. Avšak vzhledem k tomu, že používané váhy jsou ve všech těchto sítích stejné, jejich množství je maximálně n^2 . Důvod, proč tedy dropout funguje, je následující. Největší generalizace by bylo totiž docíleno, pokud by bylo natréno-váno velké množství neuronových sítí a jejich výstupy by byly zprůměrovány. Toto ale není reálně příliš proveditelné. Místo toho se toto řešení approximuje. A právě dropout je touto approximací. Při dropoutu totiž dochází k trénování velkého množství sítí avšak se stejnými parametry [41].

V rekurentní neuronové síti je ale implementace dropoutu rozdílná – dropout je možné provést více způsoby. Lze vypínat rekurentní spojení, spojení s další vrstvou nebo obě tato spojení. Způsob implementace dropoutu v rekurentní neuronové síti byl samozřejmě zkoumán, a to v [34], [49] a dalších pracích. V obou případech byl zvolen způsob, kdy dochází k vypínání spojení skryté vrstvy s další vrstvou (viz 3.15). Tento způsob bude tedy použit i v této práci.



Obrázek 3.14: Vlevo běžná neuronová síť, vpravo síť s aplikací dropout [41]



Obrázek 3.15: Dropout v rekurentní neuronové síti [34]

Kapitola 4

Data

V této kapitole budou popsány jednotlivé zdroje dat. Zároveň budou data patřičným způsobem komentována, popsána a budou vyloženy jejich případné nedostatky.

4.1 Výběr produktů

Jelikož fundamentální analýza má větší smysl pro predikci v delším časovém rámci, byly za primární testovaný produkt vybrány akcie. Jejich obchodování je specifické v tom, že spekulace na růst akcií je vhodná pouze dlouhodobější. Koupit akcie a držet je pouze jeden den není nijak efektivní (díky absenci pákového efektu). Ceny akcií firem lze zároveň považovat za produkt, u kterého má fundamentální analýza smysl. Jelikož se na světových burzách obchoduje s akcemi velkého množství firem, bylo vybráno pouze 100 firem z největší burzy na světě – NYSE (New York Stock Exchange) podle délky jejich obchodování na burze (důvod bude vysvětlen dále). Seznam těchto firem je uveden v tabulce 4.1. Jako zdroj dat je použit seznam aktuálně obchodovaných firem na burze NYSE na stránkách nasdaq.com. Alternativně lze přidat forexové produkty, které není problém obchodovat v delších časových rámcích, ale na rozdíl od akcií, s neporovnatelně menším kapitálem.

U testování systému na minulých datech u akcií je potřeba si dát pozor na tzv. *survivorship bias* [38]. Problém tkví v tom, že firmy musí pro obchodování na burze plnit nějaké podmínky. Při nesplnění těchto podmínek dochází k vyřazení firmy z burzy. Nejčastěji se jedná o podmínky a důvody typu – minimální cena akcie, minimální hodnota tržní kapitalizace, bankrot apod. Na burze NYSE se jedná o cca 200 firem ročně (z cca 3000 firem) a na burze NASDAQ asi 600 firem ročně, které jsou z burzy vyřazeny [21][29][30]. Pokud vybereme firmy obchodované na burze například v roce 2015 a testovat výkon systému budeme například od roku 2010, pak už je v datech zanesena informace, která by v roce 2010 neměla být známa. O testovaných firmách v roce 2010 lze totiž říci, že s jistotou budou na burze i v roce 2015 a tedy se jim zřejmě nepovede špatně. Pro správné testování by bylo nutné vybrat firmy aktuálně obchodované v roce 2010 a například každý měsíc tento seznam firem aktualizovat. Získat data o firmách, které byly na burze obchodované k určitému datu je ale poměrně složité.

Architektura takového systému by byla řádově složitější. Navíc potřebné kvalitní zdroje dat, které by všechny potřebné informace obsahovaly, již nejsou volně veřejně zdarma dostupné. Proto byly jednoduše vybrány firmy aktuální v roce 2015 a tento seznam firem se v průběhu testování neměnil. Zároveň byl v této práci zvolen postup, jak alespoň částečně *survivorship bias* předejít. Vybrány byly pouze nejstabilnější firmy, což mohlo spolu s vě-

domím nízkého počtu vyřazených firem z burzy ročně, alespoň částečně *survivorship bias* snížit. Za nejstabilnější firmy přitom považujeme firmy, které se na burze obchodují nejdéle. Pro opravdu kvalitní testování by přesto bylo potřeba seznam firem průběžně měnit.

4.2 Data produktů

Historická ceny akcií firem poskytuje server <http://finance.yahoo.com/>. Tento server poskytuje pouze data aktuálně obchodovaných firem, a proto by k vyhnutí se *survivorship bias* bylo potřeba hledat jiný zdroj dat. U historických denních cen akcií jsou uváděny dvě ceny *close*, tedy ceny na konci dne. První je běžná cena *close*, která uvádí cenu aktuální v daný den. Druhá cena se nazývá *adjusted close* a ta uvádí cenu, která je přepočítaná vzhledem k datu poslední známé ceny (většinou tedy data, kdy jsou data stahována). Cena akcie se totiž v některých případech mění skokově.

Firma se může rozhodnout akcie „naředit“, například pokud je cena akcie příliš vysoká. Pro příklad uvedme, že chce firma počet akcií zdvojnásobit. Držitel deseti akcií tedy po zředění vlastní akcií 20. V reakci na to se odpovídajícím způsobem upraví cena akcie – bude dvakrát menší. Druhým případem, kdy akcie mění skokově svoji hodnotu, je čas výplaty dividendy. Po výplatě dividendy cena akcie klesne obvykle právě o hodnotu dividendy. Lze to vysvětlit tak, že v době před výplatou dividendy má akcie v sobě právě hodnotu dividendy. Proto se nevyplatí akcie nakoupit těsně před výplatou dividendy.

Tyto události by vytvářely v grafu ceny akcie skokové změny, které by ale nereflektovaly skutečnost. Při zředění akcie, a tedy zmenšení její ceny, se hodnota firmy nijak nemění. Právě tyto změny reflekтуje cena *adjusted close* a tyto skokové změny nejsou tedy v této ceně obsaženy. Zároveň toto přepočítání nepřidává do dat žádnou informaci z „budoucna“. Graf ceny akcie je pořád stejný, pouze se pohybuje v jiných absolutních číslech [43].

V této práci byla data produktů (a tedy i ostatních zdrojů dat) omezena na období 3. 1. 1995 až 11. 1. 2016. Toto omezení je komentováno v sekci 5.1. Vybrané období pokrývá 7678 kalendářních dní, z toho je ale pouze 5295 dní obchodních.

4.3 Počasí

Jako zdroj počasí slouží data získaná od NOAA (National Oceanic and Atmospheric Administration), nyní vystupující jako NCEI (National Centers for Environmental Information), který vznikl sloučením tří datacenter (klimatické, geofyzikální a oceánografické) [28]. Tato instituce je zodpovědná za poskytování přístupu k datům, které se nacházejí v jednom z nejdůležitějších celosvětových datových archivů. Jejich data jsou různě členěna – poskytuje klimatická data USA, bouřková data, radarová data, klimatické mapy apod. Pro toto práci jsou ideální GHCN-D data (Global Historical Climatology Network – Daily). Jedná se o volně dostupná data až ze sta tisíc meteorologických stanic z celého světa. Tato data obsahují denní záznamy z těchto stanic a jsou denně aktualizována. To je důležité pro nasazení obchodního systému živě. Zároveň rozlišení těchto dat je denní, což je ideální, jelikož denní data budou použita i u cílových produktů.

Data z těchto meteorologických stanic (profesionální i amatérských) poskytují spoustu druhů záznamů a u každé stanice se množství těchto záznamů velmi liší. Typy klíčových dat, která záznamy nejčastěji obsahují:

- Množství srážek [mm]

- Množství sněhových srážek [mm]
- Výška sněhové pokrývky [mm]
- Maximální teplota [desetiny °C]
- Minimální teplota [desetiny °C]

Kromě těchto dat záznamy poskytují ještě informace například o tloušťce ledu na vodě, množství oblačnosti, rychlosti větru, směru větru, výšce zmrzlé půdy, době slunečního svitu, průměrné teplotě atd. Samozřejmostí je informace o poloze meteorologické stanice (zeměpisná šířka a zeměpisná délka).

Detailní popis poskytovaných dat lze nalézt v [22].

V této práci jsou použita data maximálních teplot, minimálních teplot, množství dešťových srážek a množství sněhových srážek. Tato data jsou u meteorologických stanic nalezena nejčastěji a vzhledem k ostatním typům dat jsou často bez větších výpadků.

4.4 Forex

Forexová data neboli kurzy měnových páru, jsou získaná ze serveru <https://www.quandl.com/>. Jedná se o server obsahující velké množství volně dostupných i placených databází s různorodým obsahem. Nejčastěji se jedná o databáze finančního, ekonomického a demografického charakteru. Zároveň tento server nabízí různá API pro přístup k datům, vyhledávání v nich apod. Forexová data jsou zdarma k dispozici v databázi *Wiki Exchange Rates*. V této databázi se nachází velké množství měnových páru. Jedná se o denní data, což je pro tuto práci dostačující. Data jsou na konci každého dne aktualizovány, takže při nasazení systému v živém obchodování nebude problém s tím, že by systém neměl potřebná data včas, stejně jak tomu je při vývoji systému.

Vzhledem k obchodování s akcemi firem na americké burze NYSE byly pro práci vybrány pouze měnové páry, kdy jedním z páru je americký dolar (USD).

4.5 Google Trends

Data pro popularitu vyhledávání vybraných frází na Googlu se nachází na <https://www.google.cz/trends/>. Tato data jsou pouze v relativních číslech a představují „popularitu“ vyhledávání vybrané fráze k aktuálnímu datu. Tento zdroj dat byl použit také v práci [36]. Google Trends nabízí data v různém časovém rozlišení. Pokud chceme stáhnout data maximální délky 3 měsíců, nabídne data v denním rozlišení. Pokud potřebujeme data maximální délky 1 roku, nabídne data v týdenním rozlišení. Při stahování dat delších než 1 rok, jsou k dispozici pouze data s měsíčním rozlišením. Toto pouze práci komplikuje, jelikož po stahování dat postupně po třech měsících by bylo ještě potřeba data spojit. Bohužel to není jediný problém s Google Trends. Google omezuje množství vyhledávání v Google Trends pro jednu IP adresu. K získání většího množství dat totiž nabízí placené API. Proto byl nakonec zvolen postup, kdy se stáhnou data pro celý časový úsek s měsíčním rozlišením. Jelikož systém bude pracovat s daty denního rozlišení, jsou data pro daný měsíc rozkopírována do všech dní daného měsíce. V tomto případě by určitě bylo vhodné používat sofistikovanější techniky, například chybějící data dopredikovat. Avšak pokud dataset obsahuje pouze asi 3 % dat, tak dopredikovat zbývajících 97 % nebude patrně možné. Spíše by došlo k zanesení velkého množství šumu do dat. Rozkopírování je v tomto případě vhodné,

protože nakonec budou signály převedeny do diferencí a s diferencemi s hodnotou 0 bude zacházeno stejně jak s chybějícími daty (bude rozebráno dále). Zároveň je potřeba mít data ve všech dnech, pro které známe hodnotu, jelikož některé dny budou nakonec vynechány (viz podkapitola 5.1), a proto by nakonec mohl údaj pro celý měsíc chybět. Nakonec není potřeba při převádění měsíční hodnoty na denní tuto hodnotu dělit počtem dní. Absolutní čísla nejsou důležitá, v určité fázi stejně dojde k normalizaci.

Nakonec zbývá popsat, jaká data jsou vlastně přes Google Trends sledována. V Google Trends jsou vyhledávány názvy vybraných firem. K tomu se přidává 50 nejdůležitějších klíčových slov získaných z práce [36] (viz obrázek 4.1). Vybraná slova jsou uvedena v tabulce 4.2.

4.6 WikiTrends

V tomto případě se jedná o data popisující počet zhlédnutí vybraných článků. Nejsou potřeba počty editací článků, jak naznačuje práce [25]. Tato data lze nalézt na <http://www.wikipediatrends.com/>. Data a přístup k nim je na tomto serveru naprosto volný. Jedná se o data vztažená pouze na články na anglické Wikipedii, což také odpovídá práci [25]. Data jsou k dispozici s denním rozlišením, ale pouze od roku 2008.

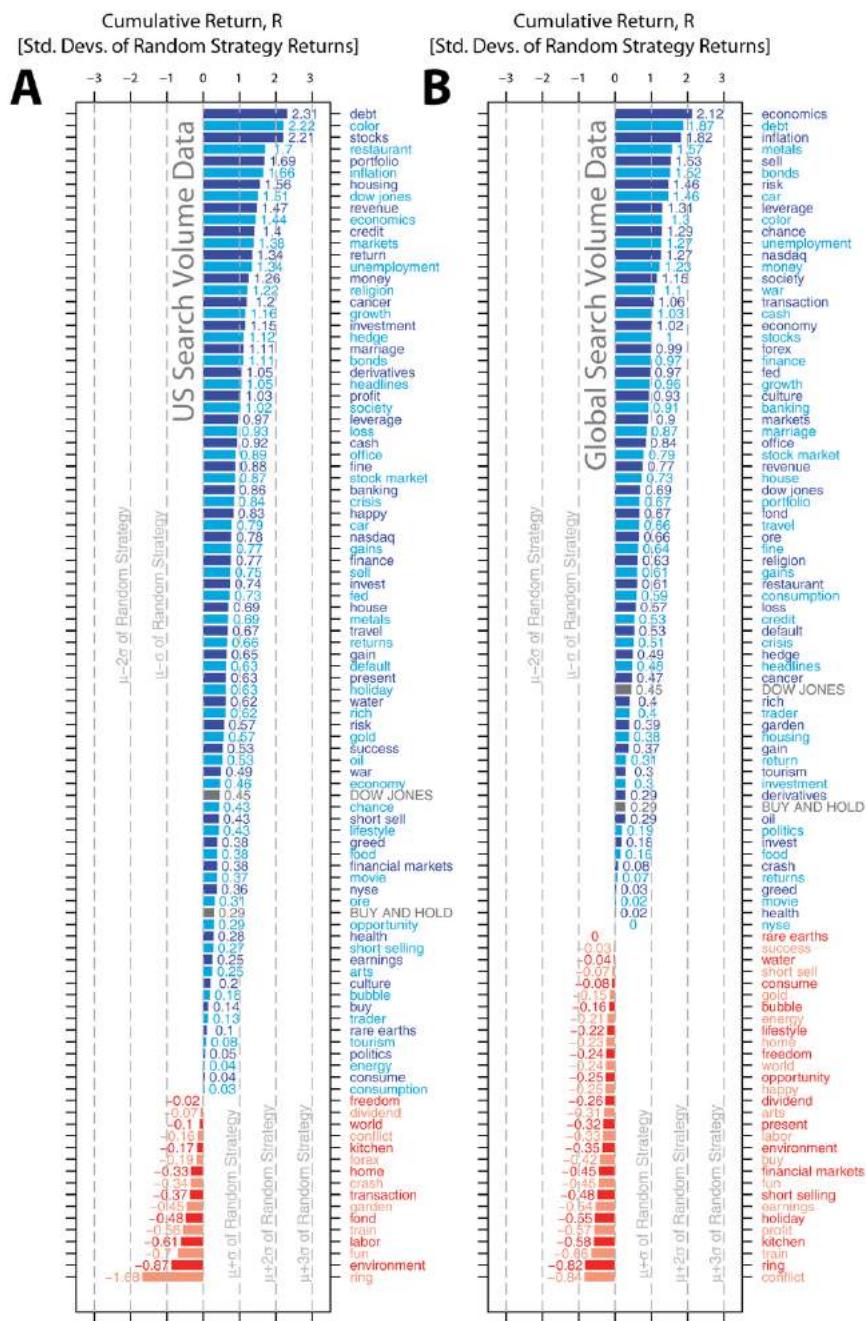
Názvy článků, které budou v této práci sledovány, odpovídají frázím hledaným v Google Trends. Jedná se tedy o články o vybraných firmách a o zvolených klíčových slovech, respektive jejich významu.

4.7 Futures

Futures data představují hodnoty kontraktů na různé komodity, finanční produkty apod. Jedná se například o zlato, stříbro, energie, kovy, státní dluhopisy atd. Získat historické ceny těchto produktů lze opět na <https://www.quandl.com/>. Konkrétní soupis všech dostupných futures je dostupný zde: <https://www.quandl.com/collections/futures>. Veškerá data se nacházejí v bezplatné databázi *Wiki Continuous Futures*. Jejich délka se velmi různí, některé futures mají dostatečně dlouhou historii (například i od roku 1950) a některé zase jsou velmi mladé (například od roku 2013). Rozlišení všech těchto dat je na denní bázi. Často se ale stává, že data pro některé dny chybí.

4.8 Fundamentals

Tato data (omezené pouze na USA) obsahují různé ekonomické ukazatele (HDP, státní rozpočty, inflace, ...), demografické ukazatele (velikost populace, úmrtnost, ...) a další (spotřeba alkoholu, ceny nájmů, délky dálnic, ...). Jejich přehled je opět na <https://www.quandl.com/>, konkrétně zde: <https://www.quandl.com/collections/usa>. Všechna tato data jsou opět bezplatně dostupná, a to v databázi *Federal Reserve Economic Data*. Na rozdíl od futures, granularita a délka těchto dat se velmi různí. Při nedostatečném rozlišení se proto podobně jak u Google Trends aplikuje rozkopírování.



Obrázek 4.1: Klíčová slova s největším významem pro predikci

Ticker	Název firmy	Ticker	Název firmy
AA	Alcoa	HON	Honeywell International
AEP	American Electric Power	HPQ	HP
AET	Aetna	CHE	Chemed
ALE	Allele	IBM	International Business Machines
APA	Apache	IP	International Paper
ASA	ASA Gold and Precious Metals Limited	JNJ	Johnson Johnson
AVP	Avon Products	KO	Coca-Cola
AVT	Avnet	KR	Kroger
AXP	American Express	LLY	Eli Lilly and Company
AXR	AMREP	LMT	Lockheed Martin
BA	Boeing	LUV	Southwest Airlines
BIO	Bio-Rad Laboratories	LXU	Lsb Industries
BK	Bank Of New York Mellon	MATX	Matson
BMY	Bristol-Myers Squibb	MCD	McDonald's
BP	BP	MDC	M.D.C. Holdings
BRT	BRT Realty Trust	MEG	Media General
C	Citigroup	MMM	3M
CAJ	Canon	MO	Altria Group
CAS	Castle (A.M.)	MRK	Merck
CAT	Caterpillar	MRO	Marathon Oil
CL	Colgate-Palmolive	MSI	Motorola Solutions
CNP	CenterPoint Energy	NAV	Navistar International
CVX	Chevron	NC	NACCO Industries
DCO	Ducommun Incorporated	NYT	New York Times
DD	E.I. du Pont de Nemours and Company	PBI	Pitney Bowes
DE	Deere	PCG	Pacific Gas Electric
DIS	Walt Disney	PEG	Public Service Enterprise Group Incorporated
DOW	Dow Chemical	PEI	Pennsylvania Real Estate Investment Trust
DTE	DTE Energy	PEP	Pepsico
ED	Consolidated Edison	PFE	Pfizer
EDE	Empire District Electric	PG	Procter Gamble
EIX	Edison International	PHI	Philippine Long Distance Telephone
EMR	Emerson Electric	PKY	Parkway Properties
ESL	Esterline Technologies	PNR	Pentair
ETN	Eaton	R	Ryder System
ETR	Entergy	SJW	SJW
EXC	Exelon	SNE	Sony Corp Ord
F	Ford Motor	SPA	Sparton
FAC	First Acceptance	SYF	Sysco
FDX	FedEx	TM	Toyota Motor Corp Ltd Ord
FL	Foot Locker	TPC	Tutor Perini
FRM	Furmanite	UIS	Unisys
FRT	Federal Realty Investment Trust	UNP	Union Pacific
FUR	Winthrop Realty Trust	UTX	United Technologies
GAS	AGL Resources	WFC	Wells Fargo
GD	General Dynamics	WFT	Weatherford International
GE	General Electric	WMT	Wal-Mart Stores
GFF	Griffon	WY	Weyerhaeuser
GTY	Getty Realty	XOM	Exxon Mobil
HAL	Halliburton	XRX	Xerox

Tabulka 4.1: Seznam vybraných firem

banking	dow jones	headlines	markets	return
bonds	economics	hedge	marriage	revenue
cancer	economy	house	metals	risk
car	fed	housing	money	sell
cash	finance	chance	nasdaq	society
color	fine	inflation	office	stock market
credit	forex	invest	portfolio	stocks
crisis	gains	investment	profit	transaction
debt	growth	leverage	religion	unemployment
derivatives	happy	loss	restaurant	war

Tabuľka 4.2: Seznam vybraných slov

Kapitola 5

Předzpracování dat

V této kapitole budou popsány potřebné transformace a operace daty, které je vhodné provést před použitím v neuronové síti.

5.1 Čištění

Jsou prováděny dvě úrovně čištění dat. Nejprve se čistí data v jednotlivých zdrojích dat způsobem specifickým pro zpracovávaná data. Druhá čistící fáze nastává po zpracování všech dat a provádí se nad všemi daty bez rozlišování jednotlivých zdrojů dat. Nejprve bude popsáno čištění všech dat a až poté bude popsáno čištění jednotlivých druhů dat.

5.1.1 Globální čištění

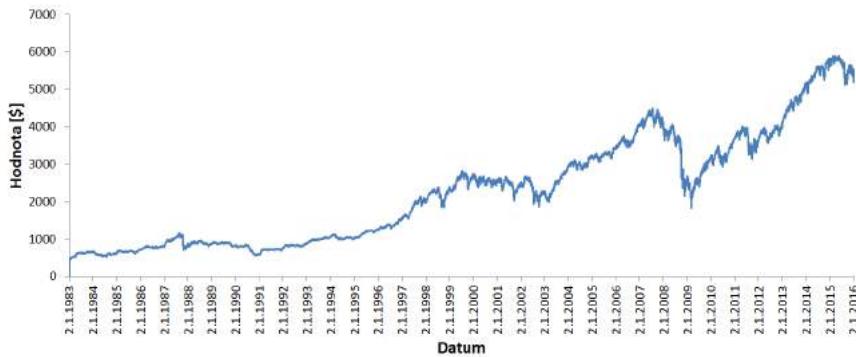
Jakmile jsou dostupná data ze všech zdrojů dat, provádí se čištění celého datasetu. Ten je uspořádán ve formě 2D matic, kde sloupce představují jednotlivé datové řady a řádky představují jednotlivé dny (čas) Nejprve v ní dochází k čištění řádků, tedy odstranění dat pro určitá data. Odstraňují se:

- Víkendy – O víkendech se s akcemi ani ostatními finančními produkty neobchoduje. Odstraněním víkendu se tak zbavíme chybějících dat o víkendech.
- Svátky – Jedná se (podobně jak u víkendů) o dny, kdy se neobchodovalo. To se určí jednoduše tak, že v ten den data všech akcií chybí.
- Příliš velká minulost – Hodnoty akcií všech 100 firem jsou dostupná nejdříve v roce 1983. Jejich cena ale ze začátku rostla velmi pomalu (viz obrázek 5.1). Proto se nakonec odstraňují všechny řádky s rokem menším než 1995.

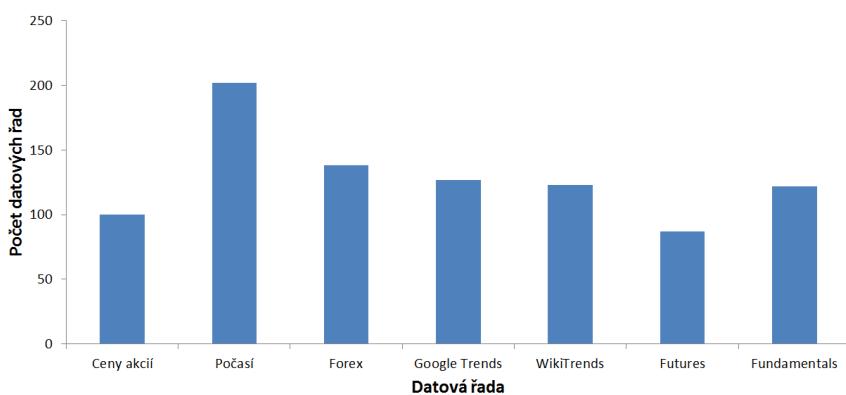
Při čištění sloupců se hledí hlavně na množství dat. V tomto případě se bere v úvahu pouze počet řádků odpovídající množství vektorů v trénovací množině. Pokud totiž v datové řadě nebude žádná hodnota v časovém intervalu určeném pro trénovací data, nemá smysl takovou datovou řadu ponechávat. Pro každý sloupec se tedy vezme pouze určitý počet řádků a zkонтroluje se, jestli neobsahuje pouze stejné nebo žádné hodnoty. Sloupce obsahující pouze jednu hodnotu (nebo žádnou) jsou odstraněny.

Přehled konečného množství datových řád pro jednotlivé zdroje dat zobrazen na obrázku 5.2. Všechna takto vyčištěná se poté transformují do podoby vhodné pro neuronovou síť.

Nyní následuje popis čištění specifický pro různé typy dat.



Obrázek 5.1: Hodnota akcií portfolia složeného ze 100 vybraných firem



Obrázek 5.2: Množství datových řad jednotlivých zdrojů dat

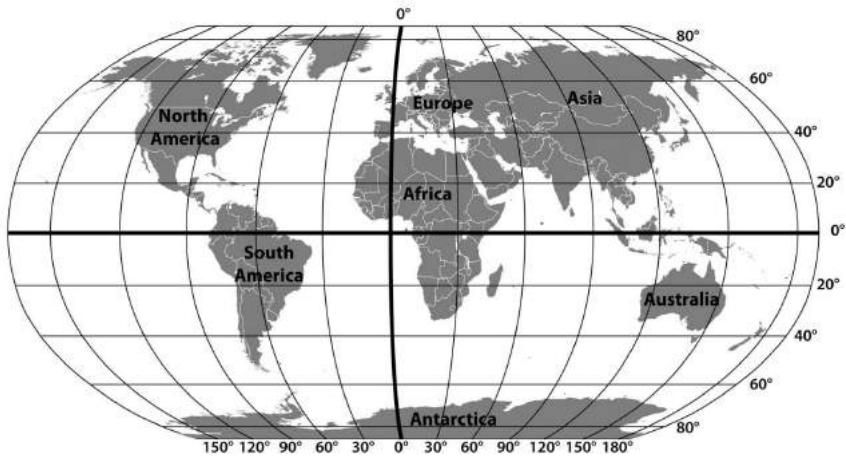
5.1.2 Akcie

Počet firem, které se budou testovat, je omezen na 100. Aby se předešlo *survivorship bias* (viz podkapitola 4.1), je vybráno 100 firem, které jsou na burze obchodovány nejdéle.

5.1.3 Počasí

Jelikož dat z počasí je velké množství (až 100000 meteorologických stanic), zahazují se všechna data mimo maximální a minimální teploty a sněhových a dešťových srážek. Dále se zahazují příliš stará data, která nebudou využita. Data z meteorologických stanic, které už svá data nepřidávají (mrtvé) se také zahazují. Kromě toho se zahazují datové řady, kde chybí více než 10 % dat.

Posledním sítěm je geografická poloha meteorologických stanic. Mapa Země byla pokryta mřížkou podobně jak na obrázku 5.3. V této práci použitá mřížka ale pokrývá pouze kontinenty a zároveň má mřížka každého kontinentu (nebo části kontinentu) jinou hustotou. Tedy někde je mřížka hustá, někde řídká a někde žádná. Při čištění dat se pro každou meteorologickou stanici nalezne nejbližší bod mřížky. Takto je u každého bodu mřížky seznam meteorologických stanic. Z těchto stanic se vyberou data jediné stanice, a to té, která má dat nejvíce. Výsledkem jsou tedy data z tolika meteorologických stanic, kolik je bodů pomyslné mřížky.



Obrázek 5.3: Pokrytí Země mřížkou [37]

5.1.4 Forex

Forexová data jsou omezena pouze na měnové páry amerického dolaru. To plyně z výběru firem obchodujících se na americké burze.

5.1.5 Futures

U futures je často více kontraktů pro jeden produkt. Tyto kontrakty mají víceméně stejné trvání, proto je vybrán pouze ten, který obsahuje nejvíce dat.

5.1.6 Ostatní

Ostatní data jsou zpracovány přesně tak, jak jsou stažena. Mezi nimi nebyly pozorovány výraznější nedostatky nebo problémy.

5.2 Chybějící data

K chybějícím datům lze přistupovat více způsoby. Zde budou rozebrány jednotlivé možnosti.

1. Nejvhodnější je chybějící data dopredikovat pomocí existujících matematických metod. Lze využít například Expectation-Maximization nebo Multiple Imputation. Jejich použití je vhodné ale pouze do určitého množství chybějících dat. Například pro Multiple Imputation je touto hranicí cca 20 % [13]. Toto je asi základní omezení pro použití těchto statistických metod.

Jako zástupce těchto metod vyberme Multiple Imputation. Při jejím použití dochází k doplnění chybějících dat různými způsoby. Doplnění chybějících dat je provedeno až n -krát. Výsledkem je tedy až n verzí kompletních dat. Na každou takovou verzi jsou pak uplatněny statistické analýzy. Následně jsou tyto analýzy zkombinovány a výsledkem je celková analýza dat. Podle této analýzy jsou potom adekvátně zkombinovány všechny verze kompletních dat do jediné výsledné [47][48].

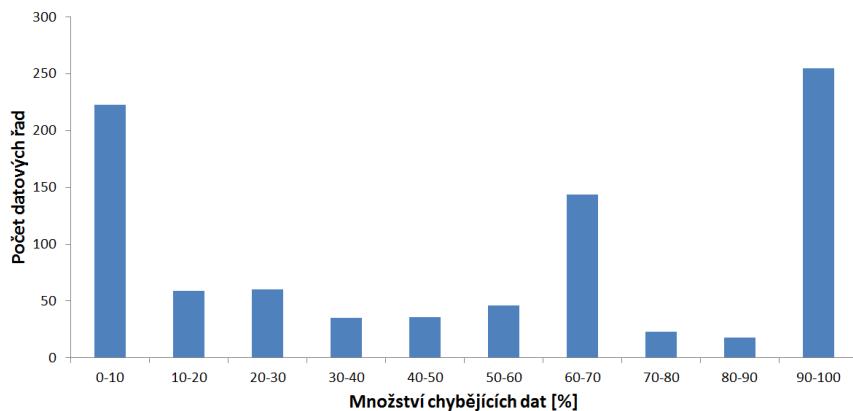
2. Chybějící data lze dále dopočítat například approximací. Okolí chybějících dat může být approximováno přímkou nebo složitější polynomiální křivkou. Limitem této metody

je nutnost znalosti sousedních hodnot a množství souvisle chybějících hodnot. Takto doplněná data také nemusí respektovat charakter dat, to už ale závisí na typu dat.

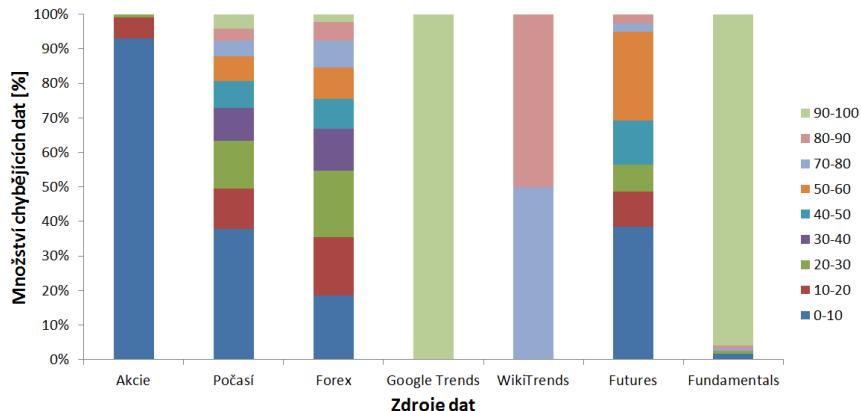
3. Namísto doplňování chybějících hodnot, mohou být tyto hodnoty nahrazeny nulami. Neuronová síť se pak může být schopna naučit tuto hodnotu ignorovat. Zároveň nula na vstupu neuronové sítě znamená to, že tato hodnota se nikak nepromítne v připojených neuronech ve skryté vrstvě. To plyne ze vzorce 3.1.
4. Místo dopočítávání hodnot lze hodnoty rozkopírovat. Tedy poslední známá hodnota se rozkopíruje namísto následujících souvisle chybějících hodnot.
5. Dále lze chybějící data generovat náhodně se stejným pravděpodobnostním rozložením jako mají známá data. Od této metody se očekává pouze jakési zaplnění prázdných vstupů. Může ale vést k tomu, že neuronová síť bude takovýto vstup ignorovat.
6. Další metodou je vypínat vstupní neurony, kterým chybí hodnota. Toto lze provádět stejným způsobem, jakým se provádí dropout, který se používá ke snížení přeúčení neuronové sítě [41]. Při vypnutí vstupních neuronů se pak hodnoty ostatních neuronů vynásobí hodnotou odpovídající poměru aktivních a neaktivních neuronů. Opět se jedná o princip použitý v dropout technice.

Všechny zmíněné metody budou dále okomentovány a některé vyzkoušeny v kapitole 8.

Pokud provedeme analýzu dat (obrázek 5.4) zjistíme, že cca třetině dat chybí do 20 % dat. Z toho plyne, že jen této části by šlo dopočítat chybějící data. Jedná se ještě o optimistický odhad. Většině dat totiž nechybí data náhodně, ale často se jedná o případy, kdy chybí data ze začátku řady (například začínají až v roce 2003). Přitom žádná další data nebudou vyřazena, protože tím bychom mohli přijít o celé zdroje dat (například o celé WikiTrends). To lze pozorovat na obrázku 5.5. Velké množství dat chybí hlavně ve skupině dat Google Trends a Fundamentals. To je způsobeno tím, že zde nejsou často k dispozici data s denním rozlišením. Vzhledem k velkému množství chybějících dat bude vyzkoušeno více přístupů a ty porovnány (viz kapitola 8).



Obrázek 5.4: Histogram množství chybějících dat



Obrázek 5.5: Množství chybějících dat podle druhu dat

5.3 NN formát dat

Není vhodné učit neuronovou síť přímé hodnoty akcií. Ceny, kterých akcie dosáhnou, mohou být totiž v datasetu jen jednou. Proto je vhodnější neuronovou síť učit hodnoty poklesu/růstu. Například při učení cen následujícího dne z cen předchozího dne bude na vstupu hodnota poklesu/růstu za předchozí den a na výstupu bude hodnota poklesu/růstu následujícího dne. Neuronová síť tedy nebude vůbec znát absolutní hodnoty cen.

Druhou důležitou informací je rozdelení dat na trénovací, validační a testovací množinu. Jejich poměr je 0,7 : 0,15 : 0,15.

5.4 Normalizace

Při aplikaci neuronové sítě na data, je potřeba nejprve data normalizovat. Nejčastější jsou dva způsoby:

- Transformace dat, aby cílová data měla střední hodnotu 0 a směrodatnou odchylku 1 (tzv. *z-score*). Tato metoda je vhodná pro data s normálním rozložením. Při jiném rozložení je vhodné data transformovat, aby normální rozložení měla. Převod konkrétní hodnoty touto metodou se provádí následovně:

$$v' = (v - \mu) \frac{\sigma'}{\sigma} + \mu' \quad (5.1)$$

kde v je původní hodnota, v' je transformovaná hodnota, μ je původní střední hodnota dat, μ' je nová střední hodnota, σ je původní směrodatná odchylka a σ' je nová směrodatná odchylka. Pro transformaci dat do normálního rozložení $\mathcal{N}_{(0,1)}$ je tedy rovnice následující:

$$v' = (v - \mu) \frac{1}{\sigma} + 0 \quad (5.2)$$

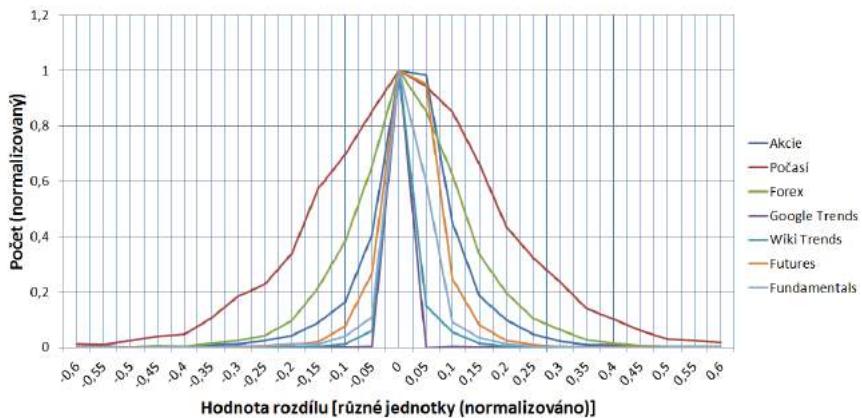
- Druhou oblíbenou metodou je lineární transformace dat do intervalu $< -1; 1 >$ (tzv. *min-max*). Metoda není vhodná v případě, kdy se mezi daty výjimečně vyskytují hodnoty řádově vyšší než většina ostatních dat. V tom případě bude většina dat ležet blízko nuly, neboť bude zastíněna vyššími hodnotami. Transformace provede následovně:

$$v' = (v - \min) \frac{\max' - \min'}{\max - \min} + \min'; \quad (5.3)$$

kde v je původní hodnota, v' je transformovaná hodnota, \max je maximální hodnota v původních datech, \max' je maximální hodnota nových dat, \min je minimální hodnota v původních datech a \min' je minimální hodnota nových dat. Pro transformaci dat do intervalu $< -1; 1 >$ je tedy vzorec následující:

$$v' = (v - \min) \frac{1 - (-1)}{\max - \min} + (-1) \quad (5.4)$$

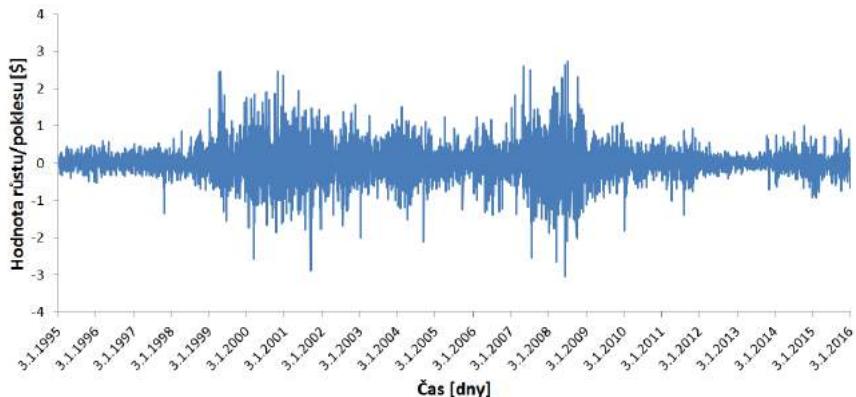
Pro zvolení vhodné normalizace je vhodné zjistit charakter vstupních dat (poklesy, růsty). Ten je vyobrazen na náhodně vybraných datových řadách z každého zdroje dat na obrázku 5.6 pomocí histogramu. Počty i hodnoty poklesu a růstu jsou normalizovány pomocí *min-max* metody kvůli vhodnému zobrazení. Z obrázku je patrné, že všechna vstupní data mají normální rozložení.



Obrázek 5.6: Rozložení vstupních dat

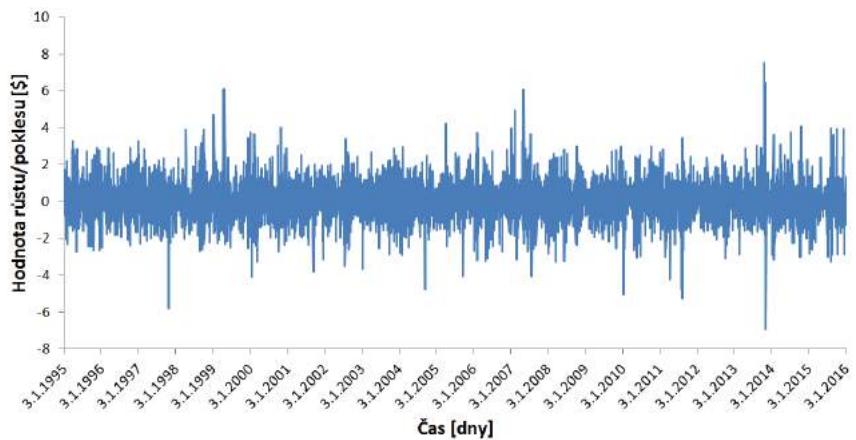
Normalizovat data klasickou metodou *z-score* by tedy neměl být problém. Pokud si zobrazíme vstupní data (přetransformována do rozdílů) v čase (viz obrázek 5.7), hned na první pohled je vidět že data jsou nestacionární. Objevují se časové volatilnější oblasti a oblasti méně volatilní. To může být při učení neuronové sítě problém. Oblasti s menší výchylkou budou způsobovat malou chybu, zatímco oblasti s velkou výchylkou budou způsobovat chybu větší. Neuronová síť se tedy bude učit z jednotlivých časových období různou měrou.

Řešením je tzv. okénková normalizace. Místo toho, aby se normalizoval celý dataset na jednou pomocí statistických informací získaných z trénovacího datasetu, normalizuje se dataset po částech. Velikost okénka se zvolí empiricky. Toto řešení bohužel způsobuje „skoky“



Obrázek 5.7: Vývoj ceny akcií AA

v místech dotyku začátku a konce okénka. To lze řešit *zjemněním* tohoto přechodu, jak je naznačeno v [31]. Z experimentů vyplývá pouze malé zlepšení oproti jednoduché okénkové normalizaci, a proto tato metoda nebude použita. Po normalizaci okénkovou funkcí jsou už data stacionární (viz obrázek 5.8).



Obrázek 5.8: Vývoj ceny akcií AA po okénkové normalizaci

5.5 Redukce dimenze

Protože je množství získaných dat vysoké (cca 900 časových řad, viz obrázek 5.2), tak takový vektor může být pro neuronovou síť příliš velký. Proto bude testována redukce dimenze dat. K tomuto bude použita metoda PCA (principal component analysis). Tato metoda se často používá k dekorelaci dat, redukci dimenze nebo hledání příznaků (feature extraction). Jejím výstupem jsou kolmé bázové vektory. Vektory udávají směr největší variability v datech. První vektor udává směr největší variability, druhý vektor udává směr druhé největší variability atd. Počet těchto vektorů odpovídá velikosti vstupních vektorů (datových řad) [24][40][51]. Při 50 % redukci dat se proto ponechá pouze prvních 50 % bázových vektorů. Následně se provede komprese dat:

$$\mathbf{D}_2 = \mathbf{B} \times \mathbf{D}_1 \quad (5.5)$$

kde \mathbf{D}_2 jsou redukovaná data, \mathbf{B} je matice bázových vektorů a \mathbf{D}_1 jsou originální data. Při zpětné transformaci dat je postup víceméně stejný:

$$\mathbf{D}_1 = \mathbf{B}^\top \times \mathbf{D}_2 \quad (5.6)$$

Je potřeba ale počítat s tím, že při redukci dimenze dat zpětnou transformací se tato data s originálními daty nebudou úplně shodovat.

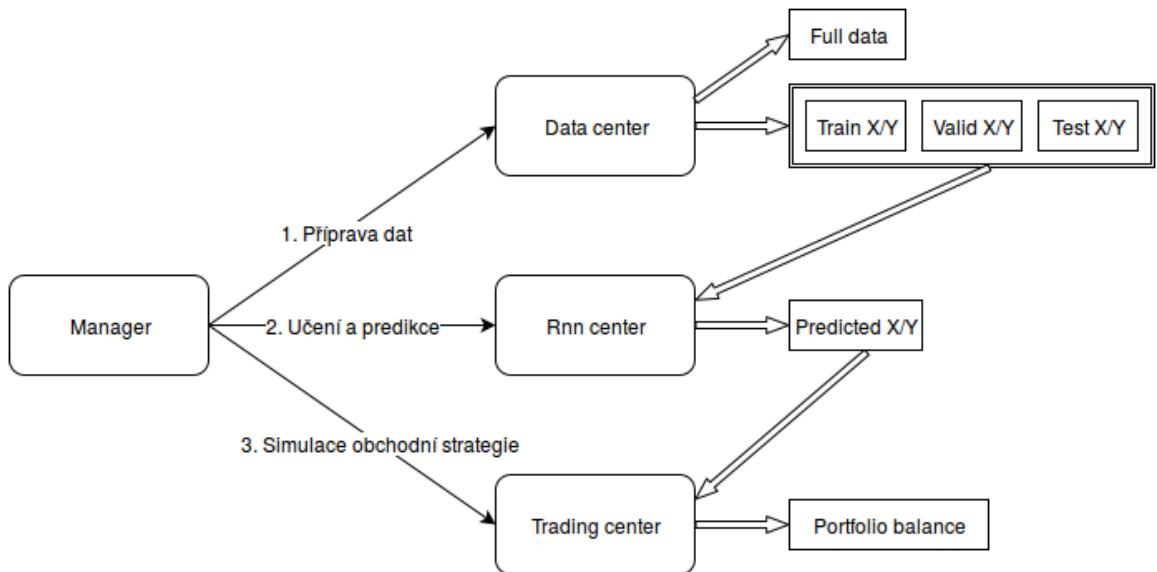
Kapitola 6

Návrh systému

V této kapitole bude popsána architektura systému, její komponenty, postup výpočtu apod. Detaily aplikace a její implementace jsou popsány v kapitole 7.

6.1 Architektura systému

Architektura celého systému je zobrazena na obrázku 6.1. Základní komponentou je *Manager*, který řídí celý proces.



Obrázek 6.1: Architektura systému

6.1.1 Data center

Prvním krokem je příprava dat. Tuto část má na starosti *Data center*. Jedná se asi o nejsložitější komponentu. *Data center* stáhne všechna požadovaná data. Kromě stažení dat data zpracuje, očistí a vybere pouze ta data, která jsou požadovaná. Takto připravená data si potom uloží, aby se celý proces nemusel opakováně provádět. Po přípravě všech druhů dat je vytvořena tabulka, kde každý řádek představuje jeden den (datum) a každý sloupec jsou

hodnoty vybrané datové řady (akcie, teploty, měny, apod.) v čase (viz ukázka v tabulce 6.1). Takto vytvořená tabulka je poté dále očištěna. Výstupem je tabulka (soubor) *Full data*.

Date	2_AA	3_NZ000093994-TMAX	4_USDGBP	5_AET	6_Economy	7_Soybeans	8_Taxes
19.5.2014	13,323096	203	0,59471	56	1546	1419,75	8,1
20.5.2014	12,968075	208	0,594319	56	1709	1415	8,1
21.5.2014	12,938491	215	0,593462	56	1510	1430,75	8,1
22.5.2014	13,046969	219	0,59268	56	1508	1452,5	8,1
23.5.2014	13,332958	220	0,593143	56	1251	1445,5	8,1
27.5.2014	13,29351	186	0,594026	56	1476	1418,75	8,1
28.5.2014	13,185032	193	0,595626	56	1443	1427	8,1
29.5.2014	13,392127	202	0,597521	56	1387	1432,75	8,3
30.5.2014	13,421712	218	0,597659	56	1325	1424,5	8,3
2.6.2014	13,638668	219	0,597175	58	2267	1427	8,3
3.6.2014	13,461158	221	0,597036	58	5212	1414,5	8,3
4.6.2014	13,628806	209	0,597317	58	1958	1416,5	8,3
5.6.2014	13,806316	209	0,596864	58	3184	1400	8,3

Tabulka 6.1: Ukázka tabulky *Full table*. Čísla sloupců značí typ dat (2 – historické ceny, 3 – počasí atd.)

Druhou funkci, kterou *Data center* zajišťuje, je příprava dat pro neuronovou síť. *Data center* načte výstupní tabulku *Full data* a vytvoří z ní šestici souborů: trainX, trainY, validX, validY, testX a testY. Jedná se o již připravená data pro neuronovou síť. Jsou tedy transformována do požadované formy a normalizována. V souborech s koncovkou *X* jsou vstupní vektory a v souborech s koncovkou *Y* jsou výstupní vektory. Vektory jsou v souborech uloženy postupně v řádcích. Každá odpovídající dvojice souborů (*X* a *Y*) má proto stejný počet řádků. Vstupnímu vektoru na řádku *n* v souboru typu *X* tedy odpovídá výstupní vektor na řádku *n* uložený v odpovídajícím souboru typu *Y*.

Jelikož *Data center* pracuje se spoustou dat, je potřeba zajistit určitou úroveň cachování, aby k časově náročným operacím nedocházelo opakováně při každém spuštění programu.

6.1.2 Rnn center

Druhou částí je naučení neuronové sítě a predikce. Tato část se provádí v *Rnn center*. Je zde využita šestice výstupních souborů s předchozího kroku. Díky takto jednoduchému souborovému rozhraní lze využít jakoukoliv implementaci neuronové sítě. Neuronová síť se naučí na trainX a trainY souborech, validace je prováděna na validX a validY souborech. Po načtenování neuronové sítě se provádí predikce na testovacích datech. Vstupními vektory jsou vektory ze souboru testX. Predikované vektory jsou uloženy v souboru predictedX. Správné výstupní vektory jsou uloženy v souboru predictedY. Z toho plyne, že obsah souboru testY a predictedY je stejný.

Denormalizaci predikovaných dat provádí opět *Data center*. Tato skutečnost není na schématu vyznačena.

6.1.3 Trading center

Posledním krokem je simulace obchodního systému. Tato část se provádí v *Trading center*. K základní simulaci postačí dvojice výstupních souborů predictedX a predictedY. Podle predikce (vektor ze souboru predictedX) bude rozhodnuto o způsobu provedení obchodu. Opravdu uskutečněný pohyb trhu pro zjištění výsledku obchodu se nalezne v souboru predictedY.

Kapitola 7

Implementace systému

V této kapitole bude probrána konkrétní implementace systému. Budou zde popsány specifické problémy, jejich řešení a další důležité aspekty vývoje.

7.1 Struktura systému

Celý systém je napsán v jazyce C# na platformě Microsoft Windows. Vývoj probíhal ve Visual Studiu. Základní komponenta **Manager** je naprogramována jako konzolová aplikace. Ostatní komponenty jsou dll knihovny, které **Manager** využívá.

7.2 Příprava jednotlivých zdrojů dat

V komponentě **DataCenter** se postupně prochází jednotlivé zdroje dat (Forex, počasí atd.) a data se připravují. Proto zde bude příprava každého druhu dat popsána zvláště. Nutno dodat, že po zpracování každého zdroje dat jsou výsledná data uložena v binárním formátu a nedochází tedy k opětovnému zpracování.

7.2.1 Akcie

Jako první je potřeba zjistit 100 firem obchodujících se na burze NYSE nejdelší dobu. To má na starosti třída **Products**. Ze stránky <http://www.nasdaq.com/>¹ se proto stáhne seznam firem, které se aktuálně obchodují na burze NYSE. Z tohoto seznamu vyčteme kromě názvu firmy i tzv. *ticker*, tedy symbol, pod kterým se firma na burze obchoduje. Datum, kdy se firma na burze začala obchodovat, v tomto seznamu ale není. Proto se provede stažení historických cen všech firem stejným způsobem, jaký je popsán v části 7.2.2. Z časových značek historických cen lze vyčíst, kdy se firma začala na burze obchodovat. Pak už stačí vybrat 100 firem, které se obchodují na burze nejdéle.

7.2.2 Historické ceny akcií

V tomto kroku se využijí zjištěné symboly pro vybrané firmy. Historické ceny lze pro konkrétní symbol najít na <http://finance.yahoo.com/>. Například historické ceny IBM jsou zde: <http://finance.yahoo.com/q/hp?s=IBM+Historical+Prices>. Yahoo umožňuje stáhnout data i ve formátu CSV. Takto stažená data se poté zpracují a z cen se vybere

¹<http://www.nasdaq.com/screening/companies-by-name.aspx?exchange=NYSE>

zavírací cena *adjusted close* (vysvětlení viz podkapitola 4.2). Veškerou tuto funkcionality v sobě zaobaluje třída `HistoricalPrices`.

7.2.3 Počasí

O přípravu těchto dat se stará třída `Weather`. Jak již bylo řečeno, počasí obsahuje data až ze 100000 meteorologických stanic. Tato data se stáhnou z http://www1.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd_all.tar.gz. Jedná se o jeden komprimovaný soubor velký necelé 3 GB. Tento komprimovaný soubor obsahuje soubor pro každou meteorologickou stanici zvlášť, tedy asi 100000 souborů. Z tohoto důvodu se neprovádí rozbalení souboru na disk. Tato operace je výkonově i časově příliš náročná. Místo toho se každý soubor ze zkomprimovaného souboru zpracovává zvlášť. Konkrétní soubor se vždy samostatně rozbalí do paměti a tam se zpracuje. Nedochází přitom k žádným operacím na disku.

Kromě tohoto souboru se stáhne soubor se seznamem meteorologických stanic² a jejich umístěním v zeměpisné šířce a délce. Ten je potřeba k čištění těchto dat.

7.2.4 Forex

Stahování forexových dat má na starosti třída `Forex`. Z url <https://www.quandl.com/collections/usa/usa-currency-exchange-rate> se vyberou všechny USD měnové páry. V tomto případě je nutné názvy měnových páru vyčíst přímo z HTML kódu. Jejich šestimístný název (například EURUSD, USDCAD apod.) pak stačí k nalezení historických cen měnového páru v databázi Wiki Exchange Rates. Poté dojde ke stažení CSV souboru s historickými cenami a jednoduchému zpracování zavírací ceny v každém dni.

7.2.5 Google Trends

Stahování dat z Google Trends je implementováno ve třídě `GoogleTrends`. Data pro Google Trends jsou volně dostupná na adrese (například pro IBM) <https://www.google.cz/trends/explore#q=ibm>. Google umožňuje tato data stáhnout přímo ve formátu CSV. Takové stažení dat je ale poměrně časově náročné, proto byla využita metoda `fetchComponent` (<https://www.google.cz/trends/fetchComponent>), která data doručí ve formátu JSON a pracuje podstatně rychleji. Ze staženého souboru ve formátu JSON jsou poté vyčteny hodnoty „popularity“ hledané fráze.

7.2.6 WikiTrends

Funkcionalita pro stahování WikiTrends dat se nachází ve třídě `WikiTrends`. V tomto případě je API poměrně přímočaré. Pro stažení dat slouží metoda <http://www.wikipediatrends.com/csv.php>, kde se do GET požadavku zakóduje název hledaného článku. Problém nastává tehdy, když článek se zadáným názvem neexistuje. To se často stává například u názvu vybraných firem. Článek sice existuje, ale název článku je napsán trochu odlišně od získaného názvu firmy. V tomto případě lze použít našeptávač dostupný pod adresou <http://www.wikipediatrends.com/typeahead.php>. Po zadání dotazu vrací seznam názvů nejpodobnějších článků. Z tohoto seznamu je pak vybrán hned první název, jelikož ten bývá nejrelevantnější hledanému názvu. Poté se požadavek na stažení dat opakuje s nově získaným názvem.

²<http://www1.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-stations.txt>

Takto stažená data jsou ve formátu CSV. Často ale trpí tím, že data chybí. V tomto případě se v CSV souboru nachází hodnota 0. Tyto hodnoty jsou tedy při zpracování souboru ignorovány.

7.2.7 Futures

O stahování futures dat se stará třída **Futures**. Získání těchto dat je poměrně složité. Jejich základní přehled se nachází na <https://www.quandl.com/collections/futures>. Zde ale nejsou všechna data, proto se z HTML souboru této stránky vyčtou odkazy skrývající se pod textem každé tabulky ve tvaru například *View All Energy Futures on Quandl (200 contracts)*. Tímto je získán seznam url, na kterých se nacházejí všechny dostupné futures. Takto se tedy stáhne každá HTML stránka. Na každé stránce se nalézá seznam futures pro dané odvětví³. Na těchto HTML stránkách se zpracuje tabulka se seznamem futures. Pro každé futures se stáhne HTML stránka zobrazující přehled jejich kontraktů. Pro každý futures tedy existuje několik různých kontraktů. Následně se přes funkci vyhledávání ověří, zda data pro tyto kontrakty existují. Následně se stáhnou data všech kontraktů pro každé futures z databáze Wiki Continuous Futures. Z těchto kontraktů je vybrán vždy pouze jeden a to ten, s největším množstvím dat. Ten často bývá právě také ten nejlikvidnější. Zbývá ze získaných dat pro kontrakty získat jejich cenu. Data pro kontrakty často kromě ceny obsahují ještě spoustu dalších dat⁴. Sloupec s cenou se určí podle obsahu klíčových slov *close*, *last* a *settle*. Po nalezení sloupce jsou všechny nalezené hodnoty zpracovány a uloženy.

7.2.8 Fundamentals

Získání těchto dat je implementováno ve třídě **Fundamentals**. U fundamentals je stažení dat podobné (ale jednodušší) jako u futures. Nejprve se stáhne HTML stránka s přehledem⁵. Podobně jako u futures je získán seznam url s kompletním přehledem fundamentals. Tyto url se nacházejí pod každou tabulkou ve formě odkazu s textem například *Detailed collection: Usa Economy*. Na takto získaných stránkách se nacházejí odkazy na data z dané oblasti. Data jsou zde seskupena tématicky v tabulkách. Protože těchto dat je opravdu hodně a v jedné tabulce jsou data často velmi podobná, je vždy vybrán pouze první dataset z každé tabulky. Takto jsou už získány url adresy určující konečnou adresu dat v databázi Federal Reserve Economic Data.

Zpracování takto stažených dat je ale složitější než u futures. Jelikož se jedná o data s různým časovým rozlišením, je nejprve potřeba určit časové rozlišení dat. To se získá průměrným rozdílem mezi sousedními daty a přiřazením tohoto rozdílu k příslušnému časovému rozlišení. Nakonec je každá hodnota rozkopírována v čase zpět podle zjištěného časového rozlišení. Tento postup je podobný jako u Google Trends.

7.3 Knihovny

V této kapitole budou zmíněny všechny použité externí knihovny.

³Například <https://www.quandl.com/collections/futures/energy>

⁴Často ještě high, low, volume, open interest apod.

⁵<https://www.quandl.com/collections/usa>

7.3.1 LSTM neuronová síť

Jako knihovna pro LSTM neuronovou síť byla zvolena knihovna SharpML-Recurrent. Tato knihovna byla vytvořena jako C# verze knihovny RecurrentJava od Thomas Lahore a RecurrentJs od Andrej Karpathy. Je distribuována pod licencí MIT, lze ji tedy bez problémů použít a modifikovat. Tato konkrétní knihovna byla zvolena proto, že je dodávána ve zdrojových kódech a lze do ní plně zasahovat. To umožňuje provádět libovolné experimenty i uvnitř neuronové sítě. Například bylo potřeba doimplementovat dropout, jelikož knihovna dropout implementovaný nemá. Tuto knihovnu lze stáhnout na <https://github.com/andrewfry/SharpML-Recurrent>.

7.3.2 PCA

Pro PCA byla zvolena knihovna Alglib⁶. Jedná se o knihovnu pro více matematických operací a metod. Pro využití v této práci stačí její volná verze, ta je totiž omezena pouze výkonově.

7.3.3 Multiple Imputation

Pro doplnění chybějících dat nebyla použita žádná externí C# knihovna, ale balíček pro programovací jazyk R. Konkrétně se jedná o knihovnu Amelia II⁷. Tato knihovna byla vytvořena speciálně pro doplňování chybějících dat.

7.3.4 Maticové operace

Pro počítání maticových operací byla zvolena knihovna Accord.NET. Tu lze nalézt na <http://accord-framework.net>. Jedná se o knihovnu bohatou na nejrůznější matematické operace, metody apod.

7.3.5 Ostatní

Mezi ostatní knihovny patří kromě knihoven spadajících pod .NET Framework ještě knihovna pro práci s archívy Tar-cs⁸, knihovna pro práci s JSON řetězci JSON.NET⁹ a knihovna pro zpracování HTML dokumentů CsQuery¹⁰.

⁶<http://www.alglib.net/>

⁷<http://gking.harvard.edu/amelia>

⁸<https://github.com/gintsgints/tar-cs/tree/master/tar-cs>

⁹<http://www.newtonsoft.com/json>

¹⁰<https://github.com/jamietre/CsQuery>

Kapitola 8

Experimentování s technikami

V této kapitole budou provedeny testy různých technik popsaných v předchozích kapitolách. Téměř všechny tyto testy budou prováděny na datech obsahující pouze historické ceny produktů. Vliv fundamentálních dat je vyhodnocován až v následující kapitole 9.

Data byla rozdělena na trénovací, validační a testovací v poměru 0,7 : 0,15 : 0,15. Trénovací data tedy zahrnují období 3.1.1995 až 14.10.2009, validační data zahrnují období 15.10.2009 až 27.11.2012 a testovací data zahrnují období 28.11.2012 až 11.1.2016. Vývoj systému, tedy veškeré experimentování, bude probíhat pouze na trénovacích a validačních datech. Na validačních datech bude strategie vylepšována a optimalizována. Výsledný systém bude poté otestován na testovacích datech. Validační data jsou tedy *in-sample* data a testovací data jsou *out-of-sample* data. Tímto by se mělo zabránit „přeoptimalizaci“ systému.

Vyhodnocení bude probíhat dle více metrik. Bude posuzována:

- Chyba neuronové sítě – dojde k posouzení hlavně rozdílu mezi chybou na trénovacích a validačních datech. Na tomto údaji půjde jednoduše poznat přeúčtení neuronové sítě.
- Chyba predikce růstu/poklesu – bude sledováno, v kolika procentech bude růst nebo pokles predikován správně. Tato metoda je doporučována v [46]. Pokud provedeme analýzu poklesů a růstů na vybraných 100 akcích, zjistíme údaje uvedené v tabulce 8.1. Z této tabulky plyne, že počet růstů velmi mírně převyšuje počet poklesů. Kromě toho jsou hodnoty růstů velmi mírně vyšší než hodnoty poklesů¹. Z této analýzy plyne více věci – pokud bychom všechny akcie nakoupili na začátku roku 1995 a nyní prodali, byli bychom v mírném zisku. A dále, bude-li úspěšnost predikcí výrazně vyšší než 50%, bude se jednat o ziskový systém. Tyto úvahy ale nezahrnují vliv poplatků.
- Křivka zůstatku účtu – bude definován jednoduchý obchodní systém a posuzován jeho výkon na trénovacích a validačních datech. Tento systém bude při predikci růstu simulovat nákup. Systém přitom nakoupí vždy pouze jedinou akci dané firmy². Pokud tedy neuronová síť bude predikovat růst pro akcie Microsoftu a Googlu, obchodní systém nakoupí jednu akci Microsoftu a jednu akci Googlu. Spekulace na pokles nebude prováděna. U akcií tato forma spekulace není běžná a obvykle je dostupná v závislosti na obchodované firmě a brokerovi.

¹Analýza probíhala na normalizovaných datech, aby všechny akcie měly stejnou váhu. Poklesy a růsty se myslí hodnoty poklesů a růstů mezi dvěma po sobě jdoucími dny.

²To se změní v sekci 9.7.

- Profit factor – jedná se o číslo reprezentující ziskovost obchodní strategie. Profit factor se vypočítá následovně:

$$PF = \frac{\sum profit}{\sum loss} \quad (8.1)$$

Výše uvedený vzorec tedy představuje pouze podíl zisků a ztrát. Z výpočtu plyne, že obchodní strategie s profit factorem vyšším než 1 bude zisková. Toto číslo se dá interpretovat také jako hodnota získaná za investovaný dolar. Profit factor obchodní strategie 1,98 tedy říká to, že za každý investovaný dolar strategie vydělá \$1,98. Čistý zisk je v tomto případě \$0,98. Z profit factoru můžeme odvodit i zhodnocení, zde se jedná konkrétně o zhodnocení 98%.

Při posuzování výsledků jednotlivých metrik bude kladen důraz hlavně na redukci přeúčení neuronové sítě a stabilitu obchodní strategie. V průběhu testování se často stane, že zisk vyvýjené obchodní strategie bude řádově o hodně nižší než zisk ostatních strategií. To bude ale dánou pouze tím, že obchodní strategie provádí méně obchodů. Obchodní strategie bude vlastně vybírat pouze obchody s nejvyšším potenciálem pro zisk. Vyššího zisku pak nebude problém dosáhnout nákupem vyššího počtu akcií. Z tohoto důvodu se při vývoji strategie nebene v úvahu konečný zisk, ale cílí se hlavně na dosažení co nejlepší **stability** obchodní strategie.

Počet růstů	257912
Počet poklesů	249350
Počet růstů [%]	50,84
Počet poklesů [%]	49,16
Hodnota růstů	15940,82
Hodnota poklesů	-15257,59
Průměrná hodnota růstu	0,0618
Průměrná hodnota poklesu	-0,0612

Tabulka 8.1: Analýza poklesů a růstů (po normalizaci)

Pro srovnání dosažených výsledků budou použity až tři obchodní strategie.

- První bude provádět náhodné nákupy, tedy někdy nakoupí a někdy ne.
- Druhá bude pouze nakupovat (*buy-and-hold*). Smyslem druhé strategie je vlastně výkonné porovnání s akciovými indexy (jak bylo řečeno v části 2.3.2). V tomto případě nebudeme výsledky porovnávat s konkrétním akciovým indexem, ale s vlastním akciovým indexem, který se skládá z oněch 100 vybraných akcí.
- Třetí strategie se bude rozhodovat na základě výsledků prediktoru prvního řádu. Takto predikovaná hodnota je vlastně stejná jak hodnota, podle které se prediktor rozhoduje. V tomto případě bude mít predikce dalšího dne stejnou hodnotu jako hodnota růstu nebo poklesu minulého dne.

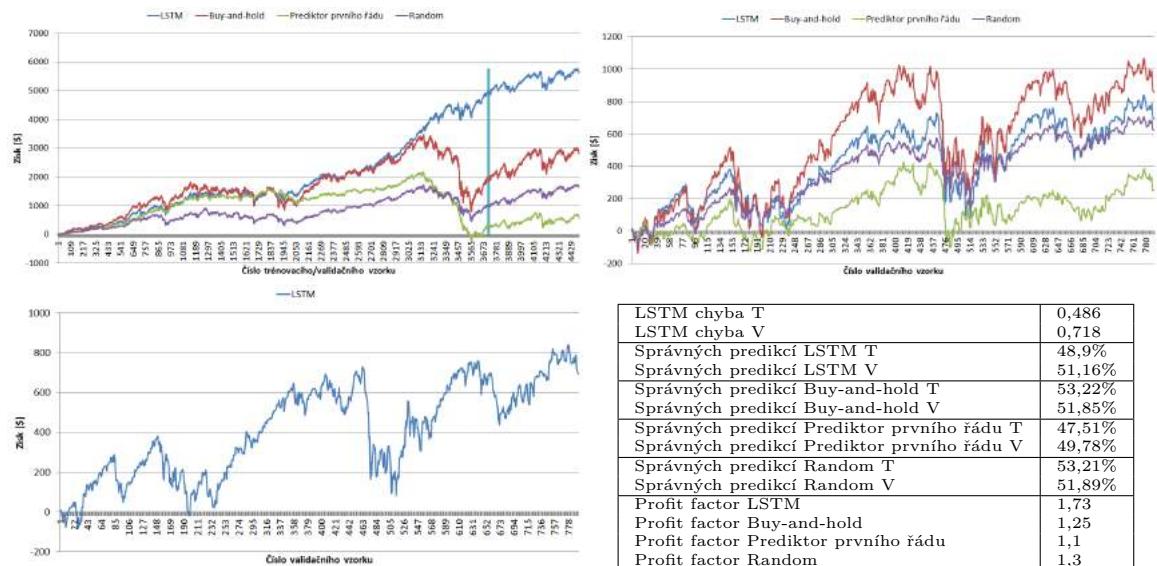
Vzhledem k tomu, že výsledky mohou být rozdílné v závislosti například na inicializaci vah neuronové sítě, každý test bude proveden vícekrát a poté vybrán reprezentativní vzorek pro daný test (cca průměrný).

8.1 Akcie – základní konfigurace

V prvním testu byla testována predikce založená pouze na hodnotách akcií. Konfiguraci lze nalézt v tabulce 8.2. Chybějící hodnoty zde nebudou řešeny, jelikož data akcií obsahují minimum chybějících hodnot. Normalizace byla provedena metodou z-score bez okénka. Statistiky pro normalizaci se samozřejmě získávají pouze z trénovacích dat. LSTM neuronová síť má jedinou skrytou vrstvu a její velikost je zde 10. Velikost vstupní a výstupní vrstvy je dána daty (zde 100). Learning rate představuje rychlosť učení. Na začátku je nastavena na nejvyšší. Při trénování se zmenší vždy, když dojde ke zvětšování chyby na validačních datech. Zároveň se načtou váhy uložené pro dosud nejmenší chybu na validačních datech. Gradient clip slouží k tomu, aby se hodnoty neuronů ve skryté vrstvě udržely v rozumných mezích. Velikost dávky určuje počet, po kolika trénovacích vektorech dojde k backpropagation. Obchodní strategie drží kupenou akcií přesně 1 den.

Výsledek této základní strategie je na 8.1. Na prvním obrázku je výkon jednotlivých strategií. Dělící úsečka odděluje trénovací a validační data. Druhý obrázek porovnává strategie pouze na validačních datech. Počátek všech strategií je zde sjednocen do nuly. Na posledním obrázku je výkon LSTM strategie na validačních datech. Oproti druhému obrázku zde jen chybí ostatní strategie. V tabulce je pak uvedena chyba neuronové sítě, počet správných predikcí v procentech a profit factor pro jednotlivé strategie. T značí trénovací data, V značí validační data. Profit factor je počítaný pouze z výsledků na validačních datech.

Z výsledků jasné plyne, že ačkoliv se neuronová síť byla schopná naučit trénovací data, došlo patrně k jejímu přeúčení. Na validačních datech je totiž obchodní strategie zhruba stejně výkonná jako ostatní porovnávané strategie.



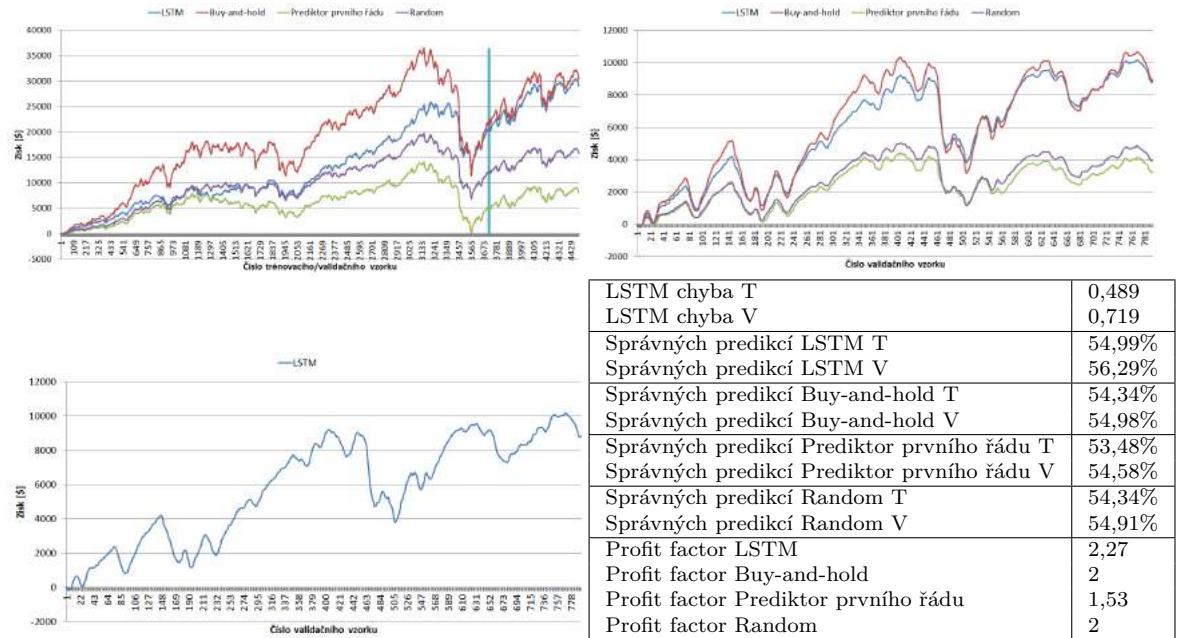
Obrázek 8.1: Test základní konfigurace

Vstupní data	historické ceny akcií
Chybějící hodnoty	rozkopírování
Vstup	pokles/růst za předchozí den
Výstup	pokles/růst dalšího dne
Normalizace	z-score, bez okénka
PCA	ne
Velikost skryté vrstvy	10
Počet iterací	40
Dropout	ne
Learning rate	0,01
Aktualizace learning rate	ano
Gradient clip	5
Velikost dávky (batch)	10
Obchodní systém – vstup	při predikci růstu
Obchodní systém – výstup	následující den

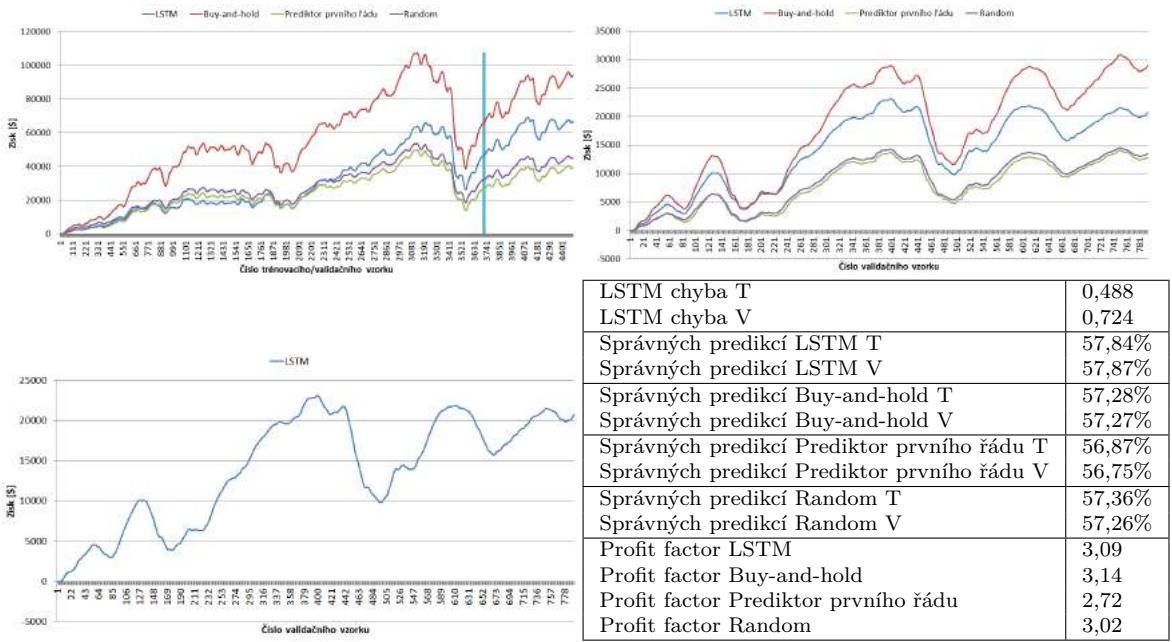
Tabulka 8.2: Počáteční konfigurace

8.2 Dlouhodobá predikce

Jako první byly vyzkoušeny dvě popisované metody dlouhodobé predikce. První testovaná metoda spočívala v opakovaném kopírování výstupu neuronové sítě na její vstup. Pro predikci n dní dopředu se tedy toto kopírování provedlo n -krát. Pokud ale predikce nefunguje spolehlivě ani na jeden den dopředu, chyba se samozřejmě se zvětšující délkom predikce zvyšuje. Pro predikci 10 dní je nejlepší výsledek zobrazen na 8.2. Přesto si strategie vede prakticky stejně jako ostatní strategie.

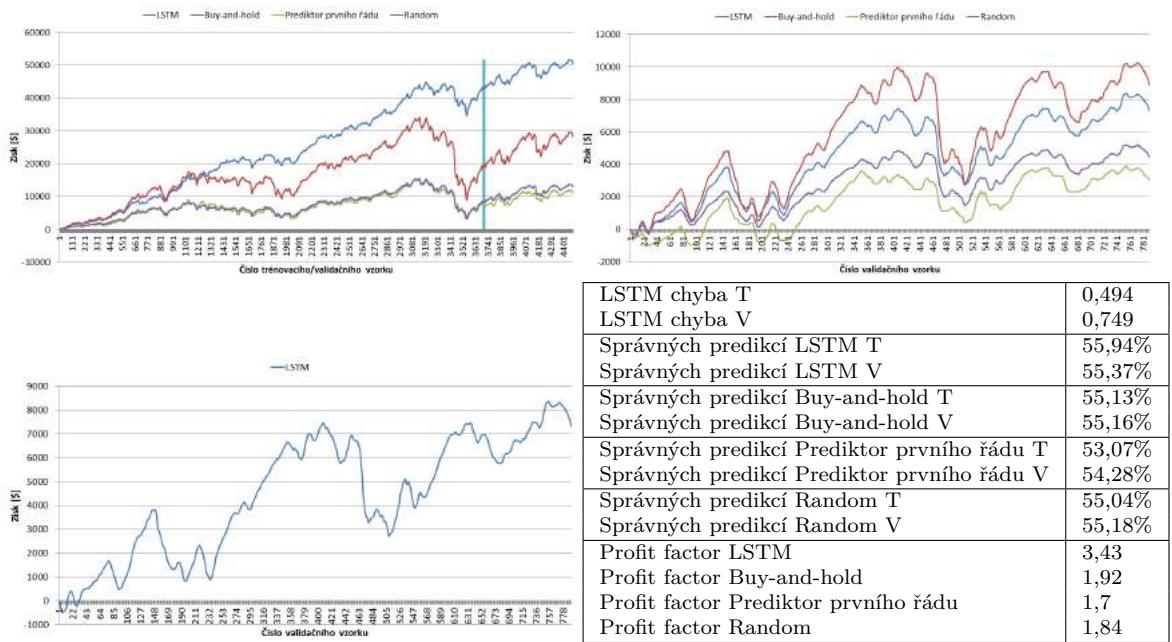


Obrázek 8.2: Test dlouhodobé predikce 10 dní kopírováním výstupu na vstup

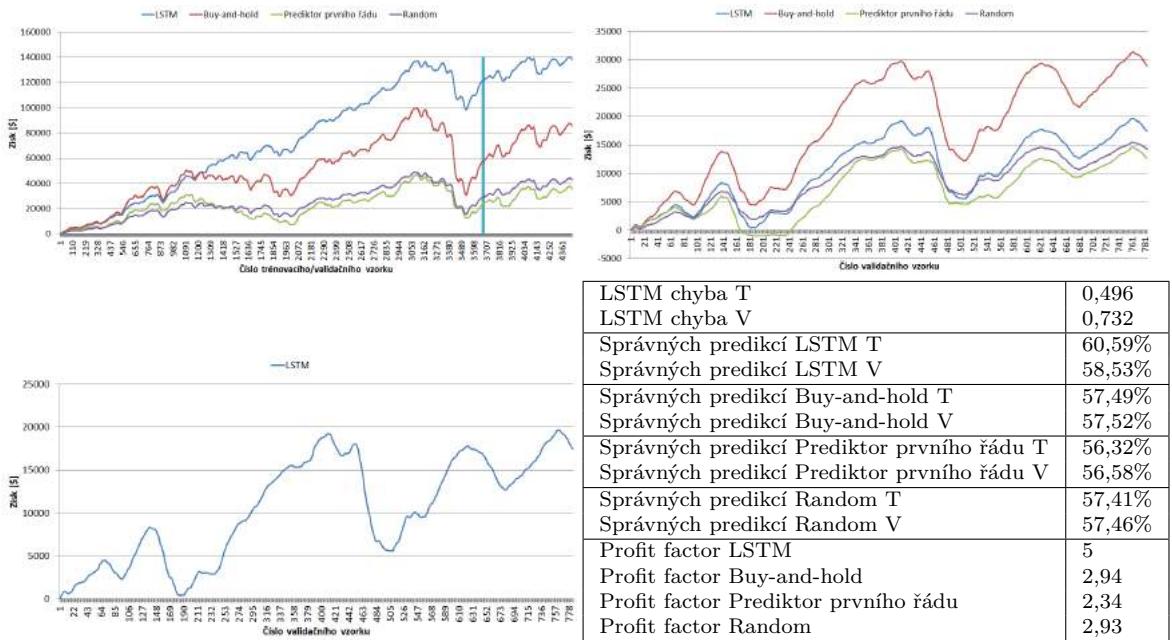


Obrázek 8.3: Test dlouhodobé predikce 30 dní kopírováním výstupu na vstup

Jako druhý způsob dlouhodobé predikce byla testována predikce růstu nebo poklesu během následujících n dní. Nejlepší výsledky pro predikci dalších 10 dní jsou zobrazeny na 8.4. Predikce pro 30 dní je zobrazena na 8.5.



Obrázek 8.4: Test přímé dlouhodobé predikce 10 dní



Obrázek 8.5: Test přímé dlouhodobé predikce 30 dní

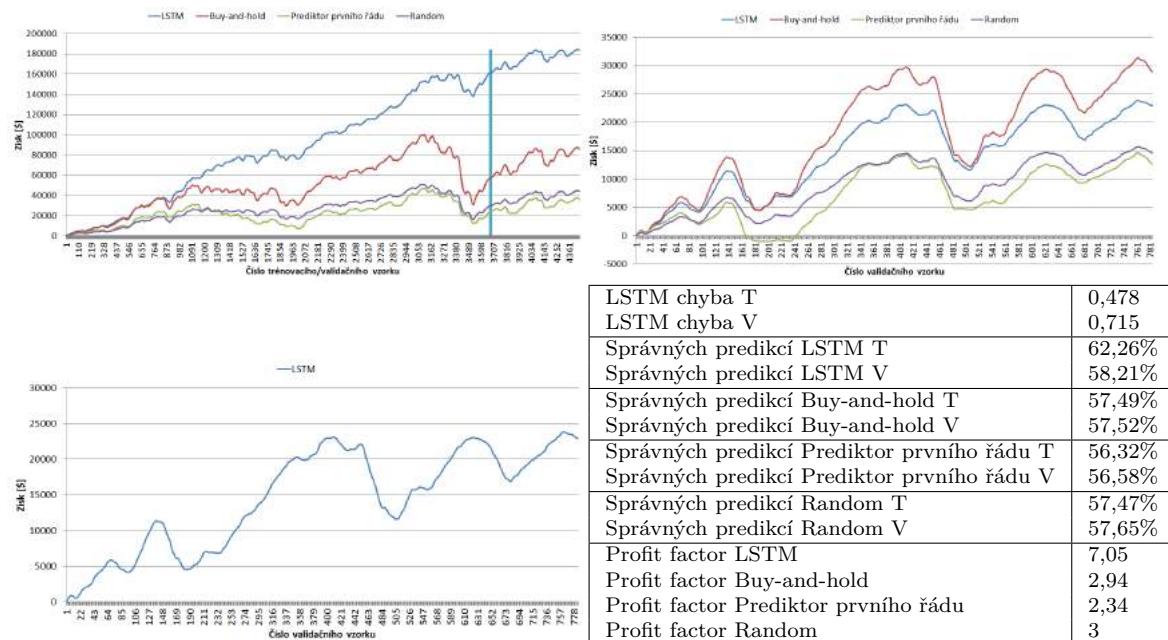
Výše uvedené testy ukazují na velmi podobné výsledky. Rozdíl je hlavně v profit factoru, který je u přímé predikce vyšší. V dalších testech bude pro dlouhodobou predikci použita přímá predikce. Tato metoda totiž s sebou nese podstatnou výhodu – data ve výstupní vrstvě mohou být prakticky libovolně transformována (různě od vstupních dat), neboť nemusí sloužit opět jako vstup neuronové sítě. Zároveň není třeba predikovat veškerá vstupní data. Výsledná konfigurace je uvedena v tabulce 8.3. Zajímavostí je skutečnost, že čím byla délka predikce delší, tím stabilnější (vyhlazenější) výsledky strategie prokazovala. Dále lze pozorovat, že výkon LSTM strategie a ostatních strategií na validačních datech je vysoce korelovaný. Liší se hlavně jejich stabilita.

Vstupní data	historické ceny akcií
Chybějící hodnoty	rozkopírování
Vstup	pokles/růst za posledních 30 dní
Výstup	pokles/růst následujících 30 dní
Normalizace	z-score, bez okénka
PCA	ne
Velikost skryté vrstvy	10
Počet iterací	40
Dropout	ne
Learning rate	0,01
Aktualizace learning rate	ano
Gradient clip	5
Velikost dávky (batch)	10
Obchodní systém – vstup	při predikci růstu
Obchodní systém – výstup	po 30 dnech

Tabulka 8.3: Konfigurace pro dlouhodobou predikci

8.3 Dropout

V této sekci bude testován dropout. Vycházet se bude z výsledné konfigurace z předchozí sekce. V této konfiguraci dropout použit nebyl, proto není třeba provádět test bez dropoutu znova. Výsledek bez dropoutu je tedy uveden na 8.5. Následuje test při zapnutém dropoutu. Dropout se aplikuje standardně, jak je doporučováno v [41]. Na vstupní vrstvě je tedy nastaven dropout 0,8 a na skryté vrstvě je dropout 0,5. Výsledek tohoto testu je zobrazen na 8.6.



Obrázek 8.6: Test aplikace dropoutu

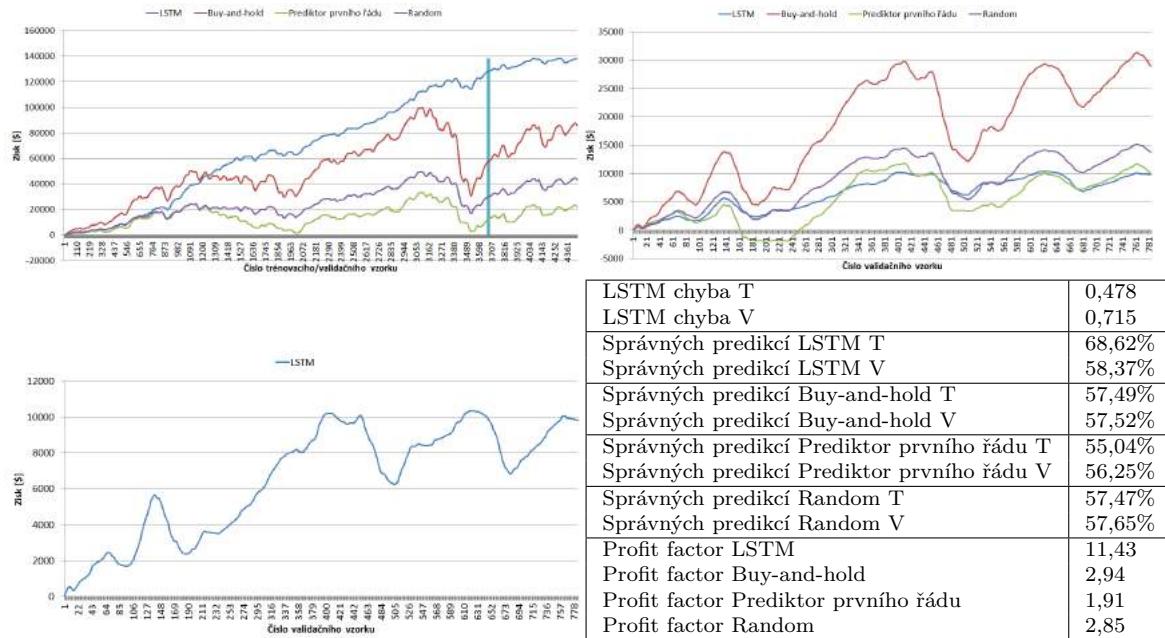
Z výsledků je patrné zlepšení výsledků LSTM strategie na validační části dat. Aplikace dropoutu také implikovala nutnost zvýšit velikost skryté vrstvy. Strategie je nyní stabilnější, mírně se zlepšila její úspěšnost i chyba na validačních datech. Přesto výkon vyvíjené strategie stále hodně koreluje s ostatními strategiemi.

Vstupní data	historické ceny akcií
Chybějící hodnoty	rozkopirování
Vstup	pokles/růst za posledních 30 dní
Výstup	pokles/růst následujících 30 dní
Normalizace	z-score, bez okénka
PCA	ne
Velikost skryté vrstvy	30
Počet iterací	40
Dropout	vstupní vrstva – 0,8, skrytá vrstva – 0,5
Learning rate	0,01
Aktualizace learning rate	ano
Gradient clip	5
Velikost dávky (batch)	10
Obchodní systém – vstup	při predikci růstu
Obchodní systém – výstup	po 30 dnech

Tabulka 8.4: Konfigurace po aplikaci dropoutu

8.4 Transformace dat

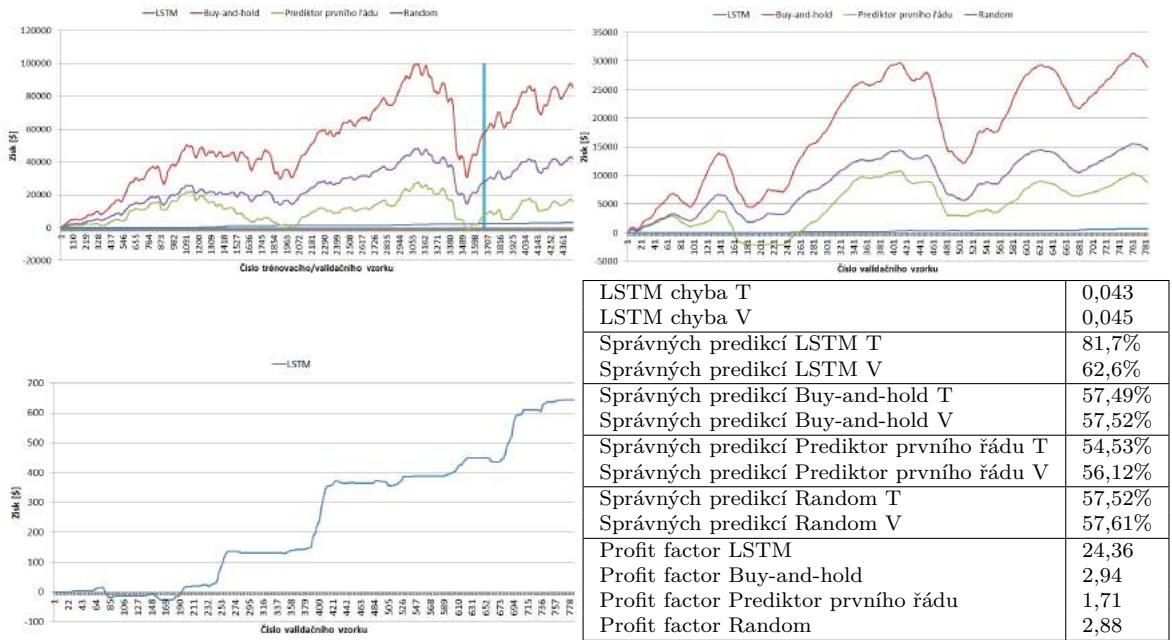
Předchozí výsledky doposud ukazují na slabý výkon neuronové sítě. Pro zlepšení výsledků lze zkoušet a používat různé přístupy. Nalezení vhodných postupů je často stěžejním bodem práce. V tomto případě se jako první nabízí kupovat pouze ty akcie, pro které není predikován pouze růst, ale pro které je predikován „dostatečně velký růst“. Velikost dostatečně velkého růstu je vždy nalezena empiricky a je definována v dolarech. Výsledek tohoto testu je zobrazen na 8.7.



Obrázek 8.7: Test hranice vstupu do obchodu v obchodním systému

Aplikace tohoto postupu očividně nepřinesla žádné výsledky. Došlo pouze ke zvýšení hodnoty profit factoru. Vhodnou volbou hranice nebylo docíleno prakticky žádného zlepšení.

Druhou možností je zakódovat výstupy neuronové sítě. Například výstup 1 bude znamenat růst, výstup 0 zase pokles. V tomto případě neuronová síť neprovádí tedy predikci, ale klasifikaci. Při testování sítě je výstup sítě samozřejmě kdekoliv v intervalu $< 0; 1 >$. Přitom čím vyšší hodnota na výstupu je, tím je predikce růstu silnější. Naopak čím blíže je výstupní hodnota nule, tím silnější je predikce poklesu. Výsledek aplikace tohoto postupu lze pozorovat na 8.8



Obrázek 8.8: Test hranice vstupu do obchodu v obchodním systému a 0-1 ve výstupu neuronové sítě

Takto získané výsledky už vypadají podstatně lépe. Profit factor oproti předchozím výsledkům výrazně vzrostl. Křivka reprezentující zisk byla výrazně vyhlazena. Došlo sice k výrazenému poklesu zisku, to je ale dánno snížením počtu obchodů. Zvýšením počtu obchodovaných akcií lze zisk libovolně zvýšit. S různě nastavenou hranicí v obchodním systému lze navíc dosáhnout zvoleného kompromisu mezi úspěšností strategie a počtem obchodů.

Tato metoda lze dále upravit a to tak, že na výstupu neuronové sítě bude 1 pouze pro „dostatečně velký“ růst. Tento přístup už ale bohužel nevedl na lepší výsledky. Nakonec se nabízí ještě zakódovat vstup stejným způsobem jako výstup. Avšak ani tato metoda nezlepšila výsledky. Dále se tedy bude pracovat se zakódovaným výstupem. Shrnutí se nalézá opět v tabulce konfigurace (8.5).

Vstupní data	historické ceny akcií
Chybějící hodnoty	rozkopírování
Vstup	pokles/růst za posledních 30 dní
Výstup	pokles/růst následujících 30 dní, zakódováno do 0-1
Normalizace	z-score, bez okénka
PCA	ne
Velikost skryté vrstvy	30
Počet iterací	40
Dropout	vstupní vrstva - 0,8, skrytá vrstva - 0,5
Learning rate	0,01
Aktualizace learning rate	ano
Gradient clip	5
Velikost dávky (batch)	10
Obchodní systém - vstup	při predikci „dostatečného“ růstu
Obchodní systém - výstup	po 30 dnech

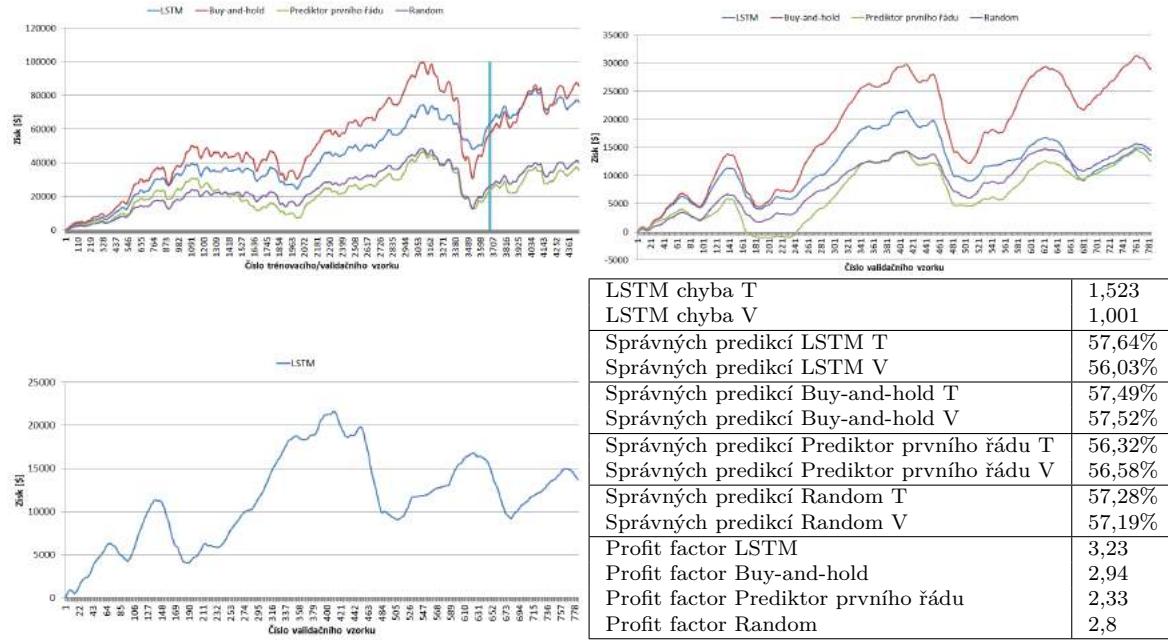
Tabulka 8.5: Konfigurace po transformaci dat

8.5 Normalizace

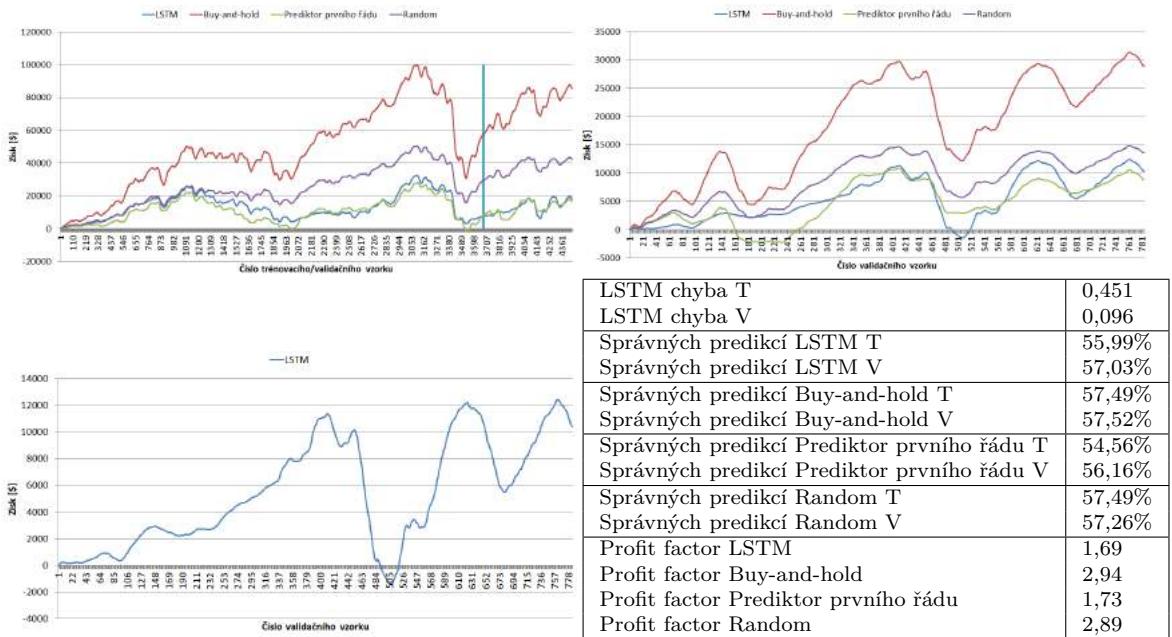
V této části bude proveden test z-score a min-max normalizace. Zároveň bude testován užitek použití okénka, jak bylo naznačeno v podkapitole 5.4.

8.5.1 Z-score

První bude proveden test z-score normalizace. Tato metoda normalizace byla až doposud používána, zbývá tedy otestovat vliv okénka. Výsledek pro okénko velikosti 50 je zobrazen na 8.9 a pro okénko velikosti 500 je zobrazen na 8.10.



Obrázek 8.9: Test z-score normalizace okénkem velikosti 50

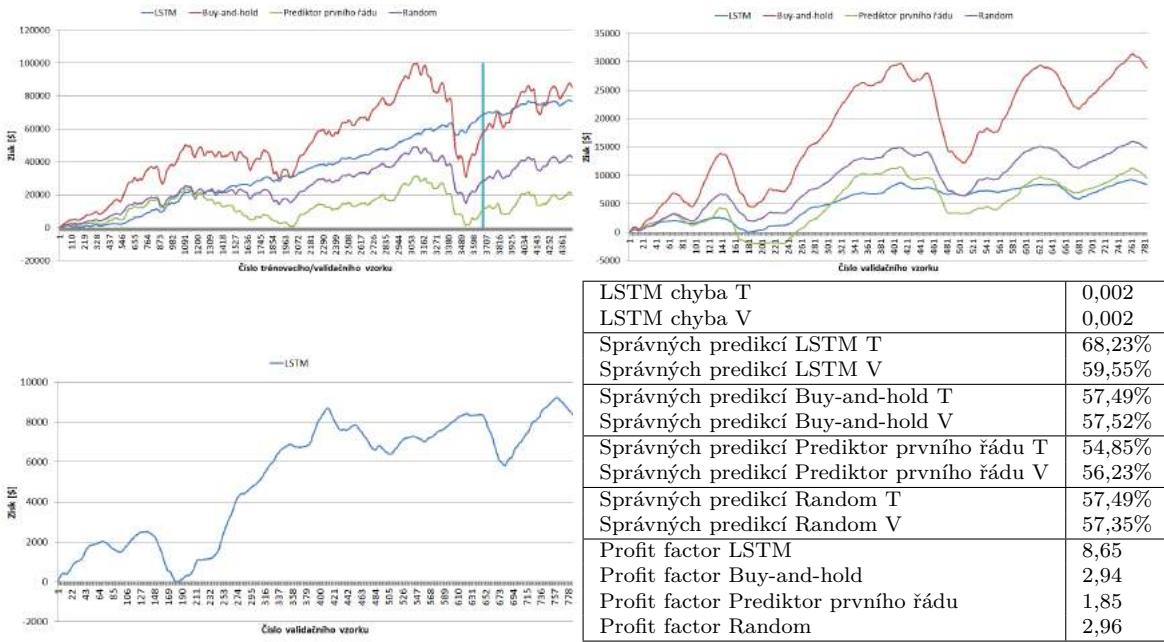


Obrázek 8.10: Test z-score normalizace okénkem velikosti 500

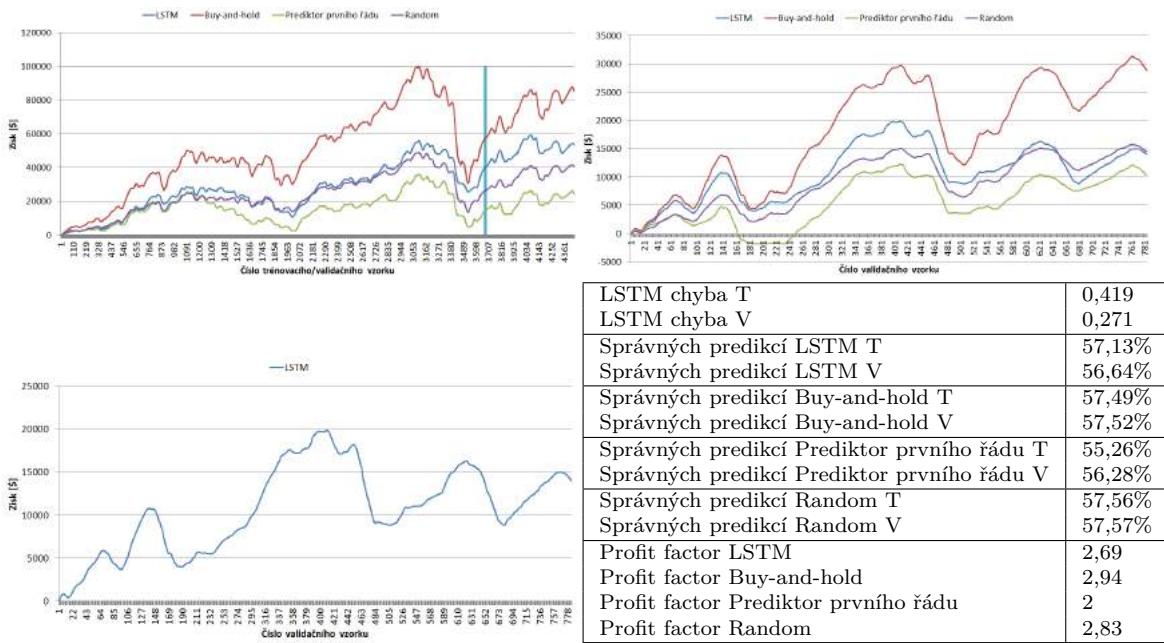
Výsledek je jednoznačně horší. Kromě nižšího výnosu je strategie méně stabilní a na validačních datech nevykazuje stálý růst. Bez použití okénka je LSTM strategie jednoznačně stabilnější. Horší úspěšnost se projevuje i na procentu úspěšných predíkcií a profit factoru. Velikost chyby neuronové sítě je také vyšší. Jednoznačným závěrem tedy je, že aplikace okénka pro metodu z-score nezlepšuje výsledky.

8.5.2 Min-max

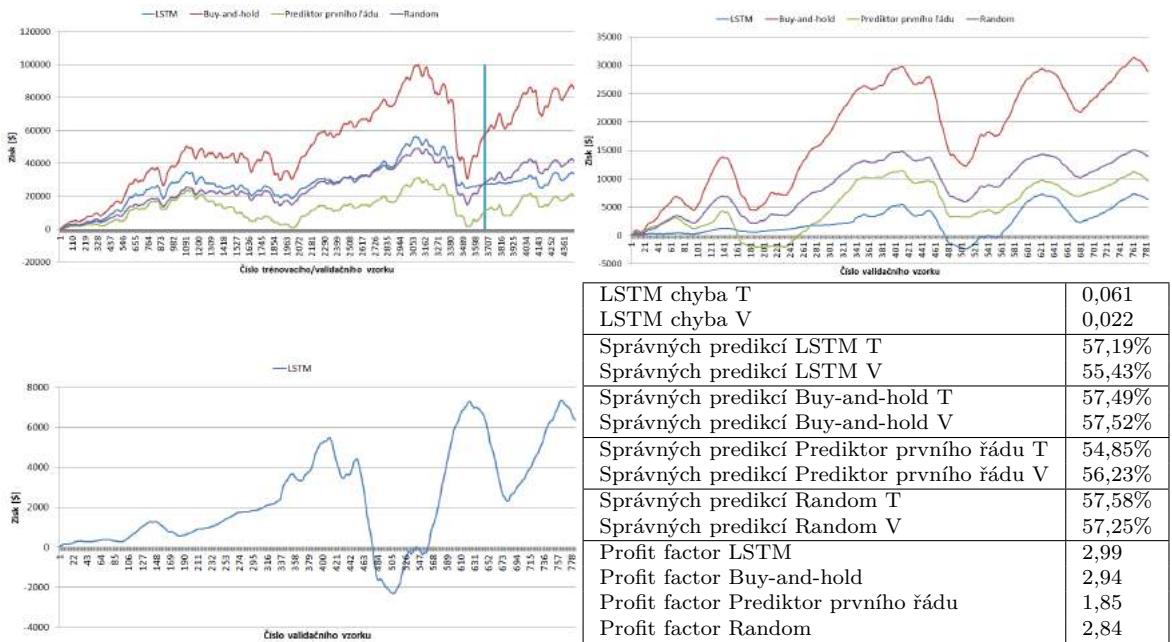
Druhou testovanou metodou normalizace je min-max. Data se tedy lineárně transformují do intervalu $< -1; 1 >$. První byl proveden test při normalizaci bez okénka (8.11) a poté opět při použití okénka 50 8.12 a 500 (8.13).



Obrázek 8.11: Test min-max normalizace bez okénka



Obrázek 8.12: Test min-max normalizace okénkem velikosti 50



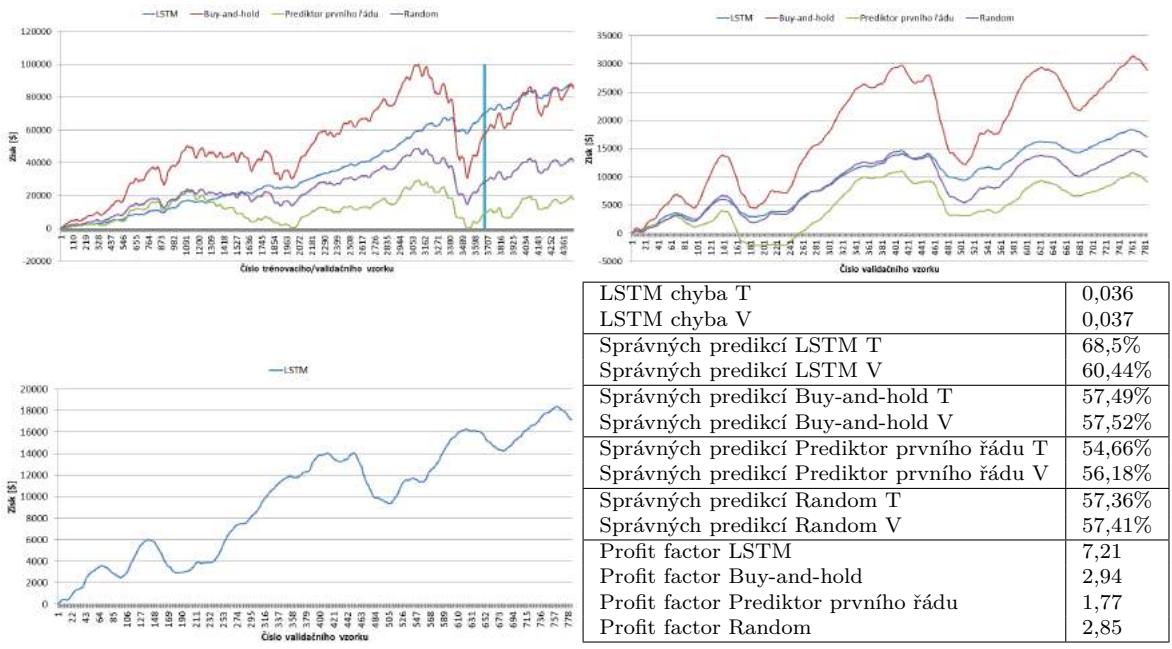
Obrázek 8.13: Test min-max normalizace okénkem velikosti 500

Z výše uvedených výsledků plyne, že použití metody min-max bez okénka sice vydělává, ale její výsledky jsou horší než při použití metody z-score. Velikost chyby neuronové sítě je nižší, ale to je způsobeno jiným charakterem dat (mají nižší hodnoty). Ani hodnota profit factoru nebyla překonána. Metody min-max s okénkem jsou na tom podobně. Nakonec ani jedna z variant nepředčila klasickou metodu normalizace – z-score. Pokračovat se proto bude se stále stejnou konfigurací, jak je uvedena v tabulce 8.5.

8.6 Chybějící data

V této části bude testován a vybrán nejlepší způsob, jak naložit s chybějícími daty. Jako testovací sada bude použita datová sada futures. Vstupní data budou tedy složena z historických cen a futures dat. Datová sada futures obsahuje střední množství chybějících dat vzhledem k ostatním datovým sadám (viz obrázek 5.5). Nebude zde hodnocen přínos futures dat, ten bude testován až později. Budou vyzkoušeny metody zmíněné v podkapitole 5.2 kromě aproximační metody. Často totiž chybí velký blok dat, například prvních 10 let apod. Na doplnění těchto dat není aproximační metoda vůbec vhodná.

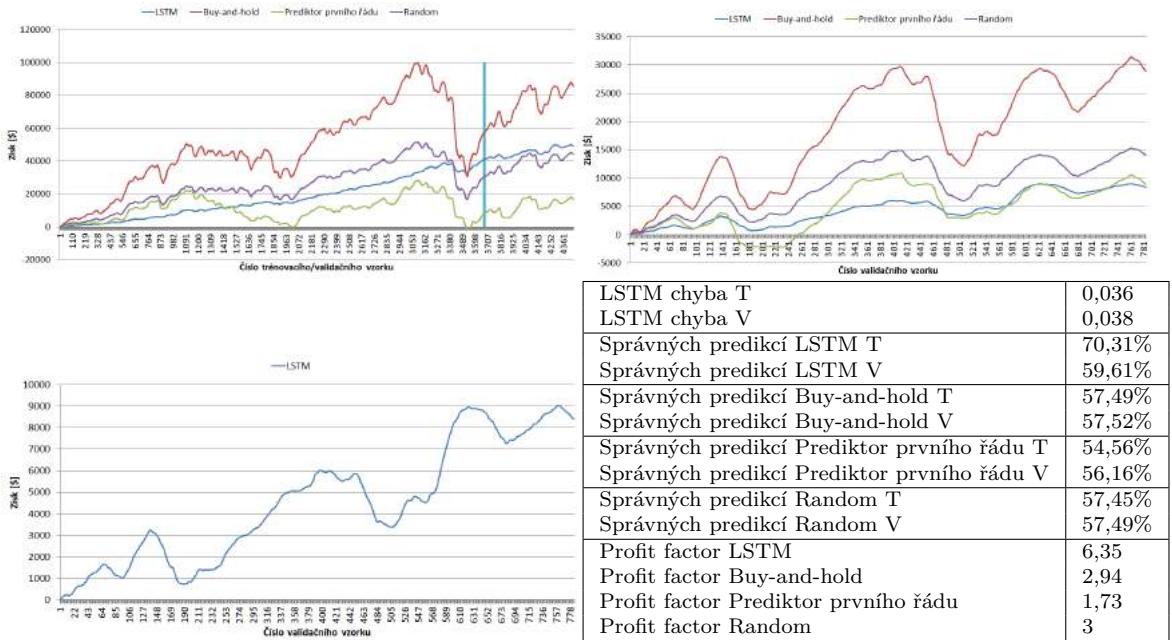
Jako první byl otestován přístup, kdy se místo chybějících hodnot dosadí nuly. Výsledek po doplnění futures dat do vstupních dat je zobrazen na 8.14.



Obrázek 8.14: Nuly místo chybějících dat

Zatím není výsledek s čím porovnávat. Říci lze zatím jen to, že při takovém použití futures dat došlo k větším chybám v predikci a strategie je méně stabilní než dosud nejlepší dosažený výsledek 8.8. Přesto se jedná o dobrý výsledek, strategie na validačních datech směřuje stále vzhůru.

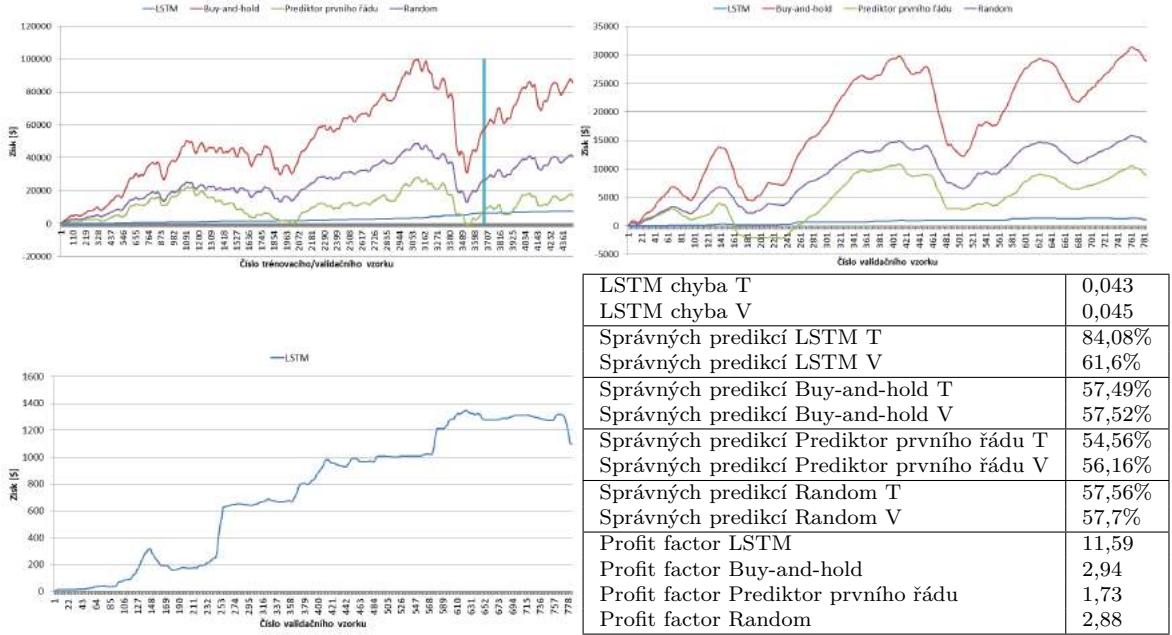
Další testovánou metodou bude vypínání vstupních neuronů s chybějícími daty. Výsledek pro tento experiment lze vidět na 8.15.



Obrázek 8.15: Vypínání vstupních neuronů s chybějícími daty

Výsledek této metody je velmi podobný předchozímu výsledku. Přesto se jedná o horší výsledek. Neuronová síť vykazuje mírně vyšší chybu, přesnost predikce i profit factor je mírně nižší a výkon strategie se na validačních datech jeví méně stabilní oproti 8.14.

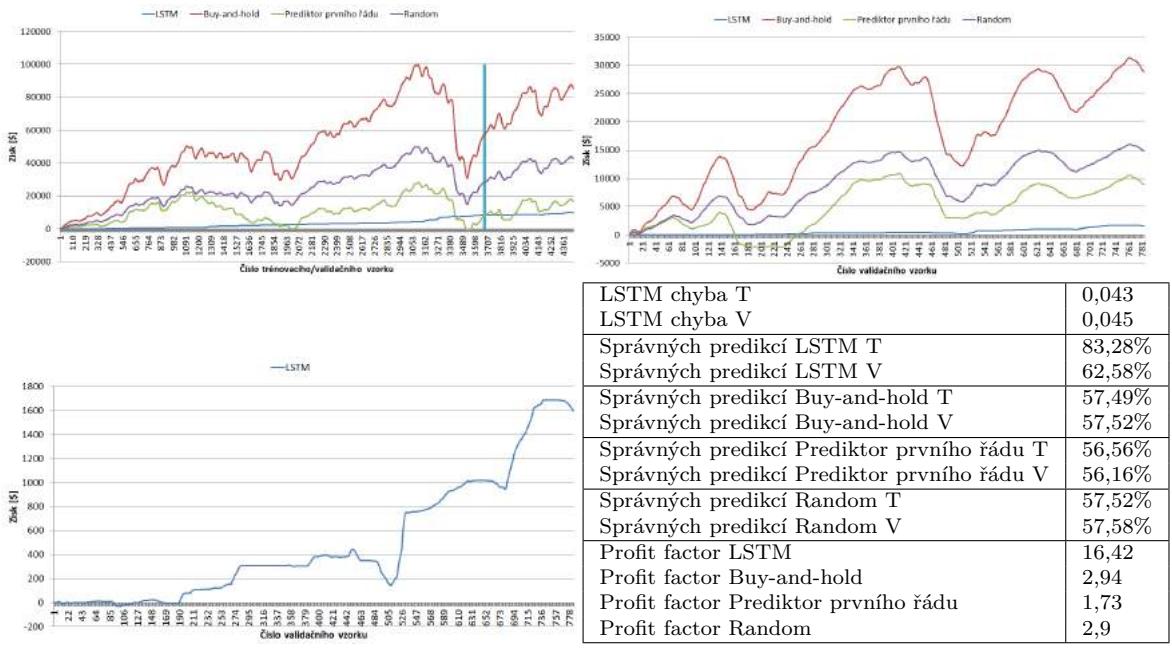
Třetím testem bude rozkopírování hodnot. Výsledek testu lze vidět na 8.16.



Obrázek 8.16: Nahrazení chybějících dat rozkopírovanými hodnotami

Tato metoda sice dosahuje nižšího profitu, ale její stabilita je podstatně vyšší. I přes vyšší chybu neuronové sítě dosahuje tato metoda vyšší úspěšnosti predikce i profit factoru a na validačních datech má strategie pouze mírné poklesy.

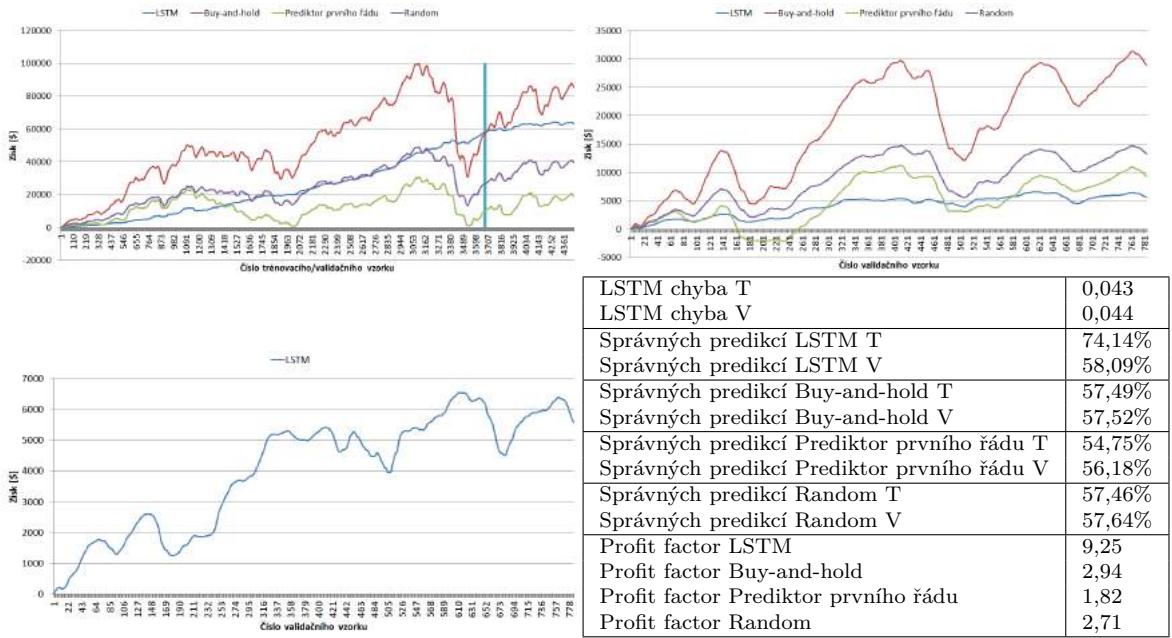
Další test bude představovat doplňování náhodných hodnot do signálu. Bude respektováno rozložení dat. Při doplňování mezer bude za střední hodnotu brána hodnota přímky approximující hodnoty v mezeře. Při chybějících datech na začátku (nebo na konci) bude za střední hodnotu brána poslední vygenerovaná hodnota. V takovém případě nebude mít vygenerovaný signál prakticky žádný směr (při střední hodnotě 0). Výsledek tohoto testu lze vidět na 8.17.



Obrázek 8.17: Nahrazení chybějících dat náhodnými hodnotami

Výsledky této metody jsou podobné s výsledkem metody rozkopírování. Avšak došlo ke zvýšení úspěšnosti predikce, i když stabilita strategie na validačních datech je poměrně podobná. Chyba neuronové sítě se nezměnila.

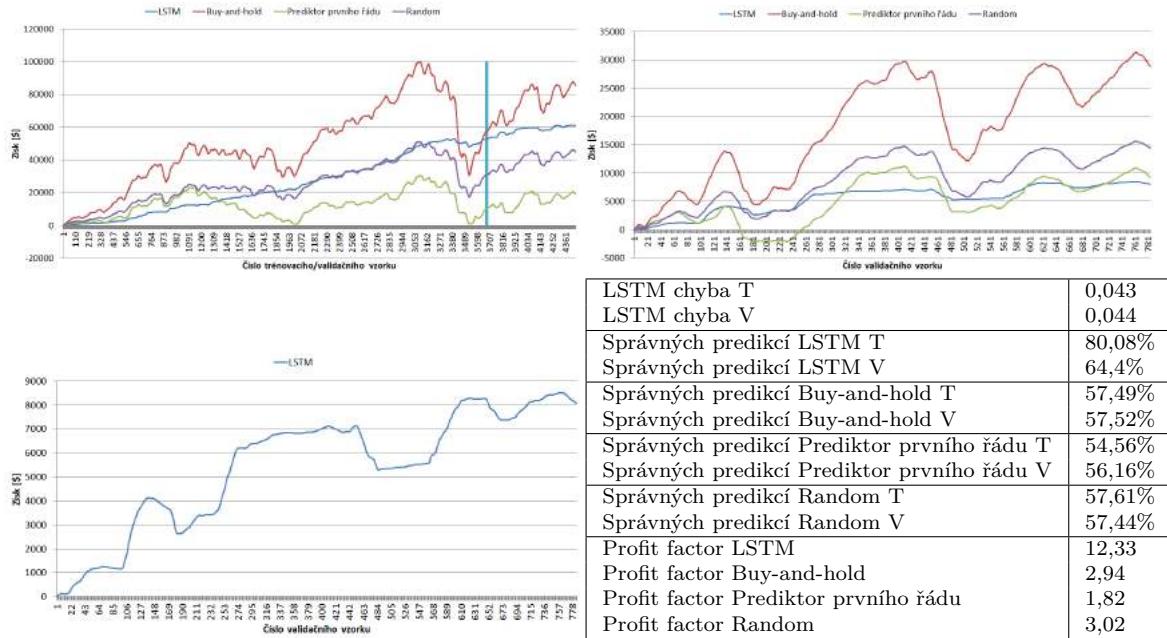
Poslední testovanou metodou bude multiple imputation. Výsledek použití této metody je zobrazen na 8.18.



Obrázek 8.18: Nahrazení chybějících dat metodou multiple imputation

Výsledek multiple imputation je jeden z těch horších. Nepříliš dobrý výsledek této metody se dal očekávat. Amelia software provádějící multiple imputation hlásil problémy

s daty. Ty plynuly hlavně z množství chybějících hodnot. Z toho důvodu nemohla tato metoda pracovat správně. U dalších zdrojů dat by byl problém ještě větší. Navíc v této metodě došlo k doplňování chybějících hodnot za znalostí hodnot budoucích (testovacích), což je pro reálné využití špatné. Avšak ani s tímto „cheatem“ tato metoda nebyla schopna překonat metodu doplňováním náhodných hodnot. Dále se tedy chybějící hodnoty budou doplňovat náhodnými hodnotami. Závěrem tohoto testování je tedy konfigurace uvedená v tabulce 8.6 a výsledek pro test systému, kde vstupní data obsahují pouze historické ceny akcií, je uveden na 8.19.



Obrázek 8.19: Test strategie s doplňováním chybějících dat

Vstupní data	historické ceny akcií
Chybějící hodnoty	náhodné hodnoty
Vstup	pokles/růst za posledních 30 dní
Výstup	pokles/růst následujících 30 dní, zakódováno do 0-1
Normalizace	z-score, bez okénka
PCA	ne
Velikost skryté vrstvy	30
Počet iterací	40
Dropout	vstupní vrstva - 0,8, skrytá vrstva - 0,5
Learning rate	0,01
Aktualizace learning rate	ano
Gradient clip	5
Velikost dávky (batch)	10
Obchodní systém - vstup	při predikci „dostatečného“ růstu po 30 dnech
Obchodní systém - výstup	

Tabulka 8.6: Konfigurace po doplňování chybějících dat

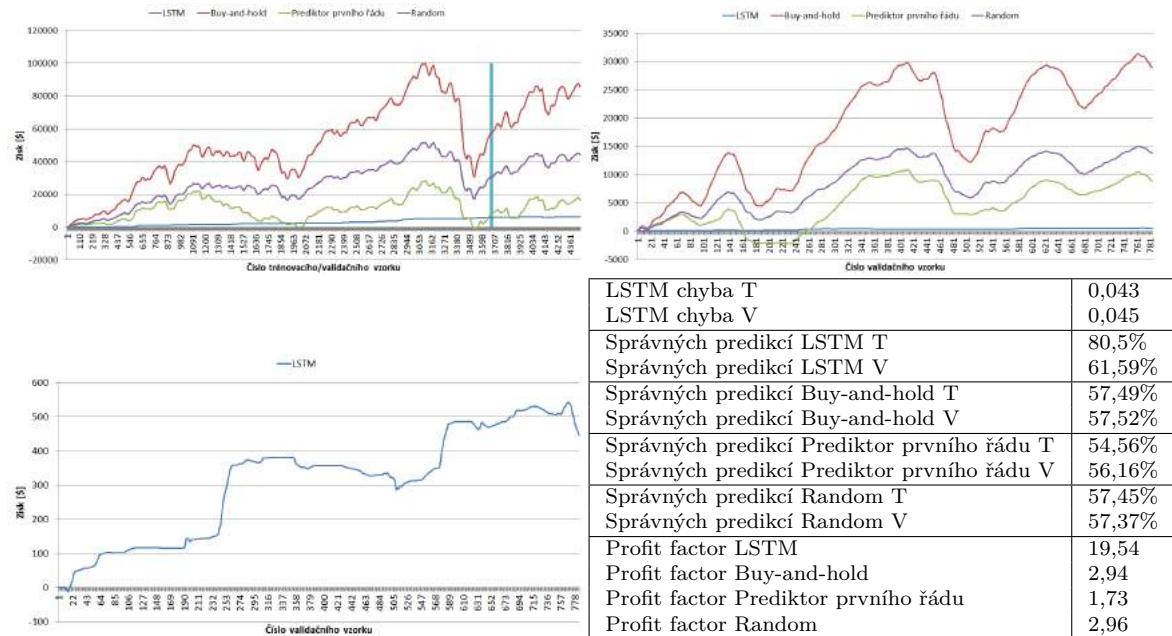
Kapitola 9

Experimentování s fundamentálními daty

Tato kapitola navazuje na předchozí kapitolu 8. Jsou zde prováděny experimenty s fundamentálními daty a dochází zde k vyhodnocení jejich vlivu na výsledky obchodního systému.

9.1 Počasí

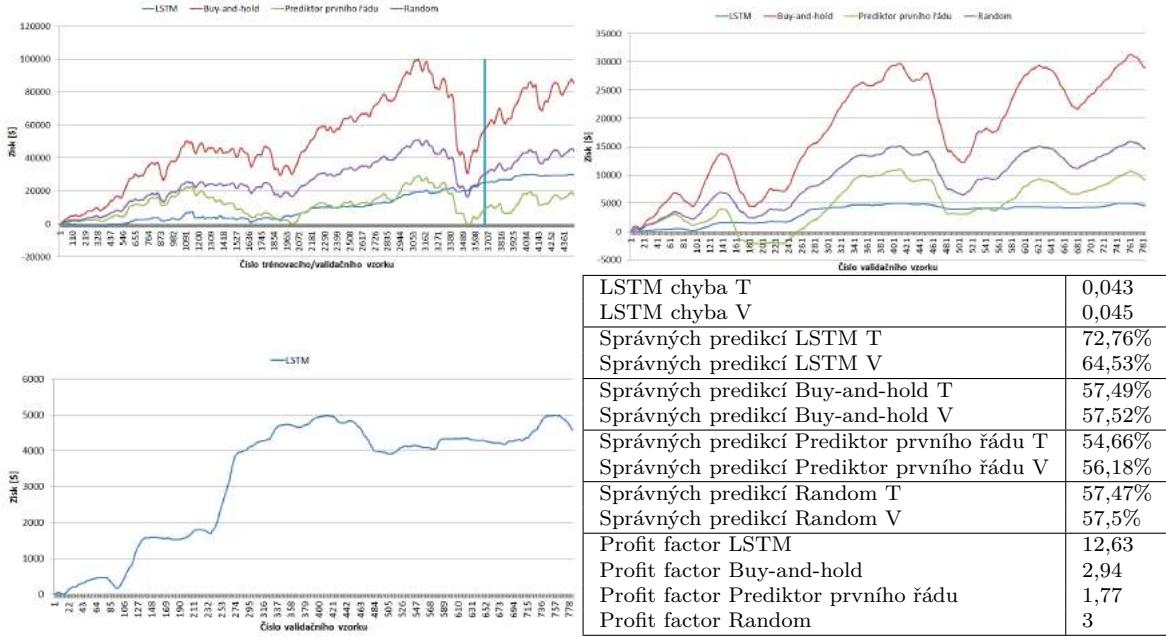
Nyní se bude postupně zkoumat vliv jednotlivých zdrojů dat na úspěšnost predikce. Jako první bude testováno počasí. Pro srovnání budeme vycházet z výsledku 8.19. U počasí nemá smysl data transformovat do rozdílů, místo toho se data za daný časový úsek sečtou. V praxi to znamená, že na vstupu neuronové sítě bude například množství dešťových srážek za posledních 30 dní. Výsledek po přidání počasí do vstupních dat je zobrazen na 9.1.



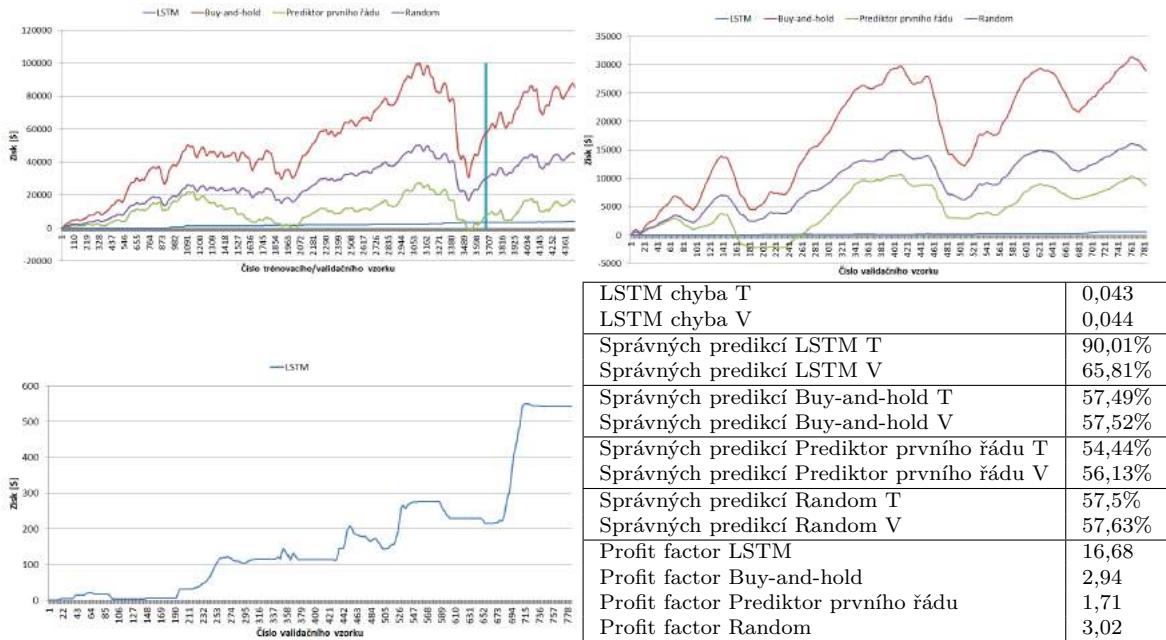
Obrázek 9.1: Přidání počasí

Při porovnání s verzí bez počasí je vidět, že přidání počasí celkové výsledky zhorsilo.

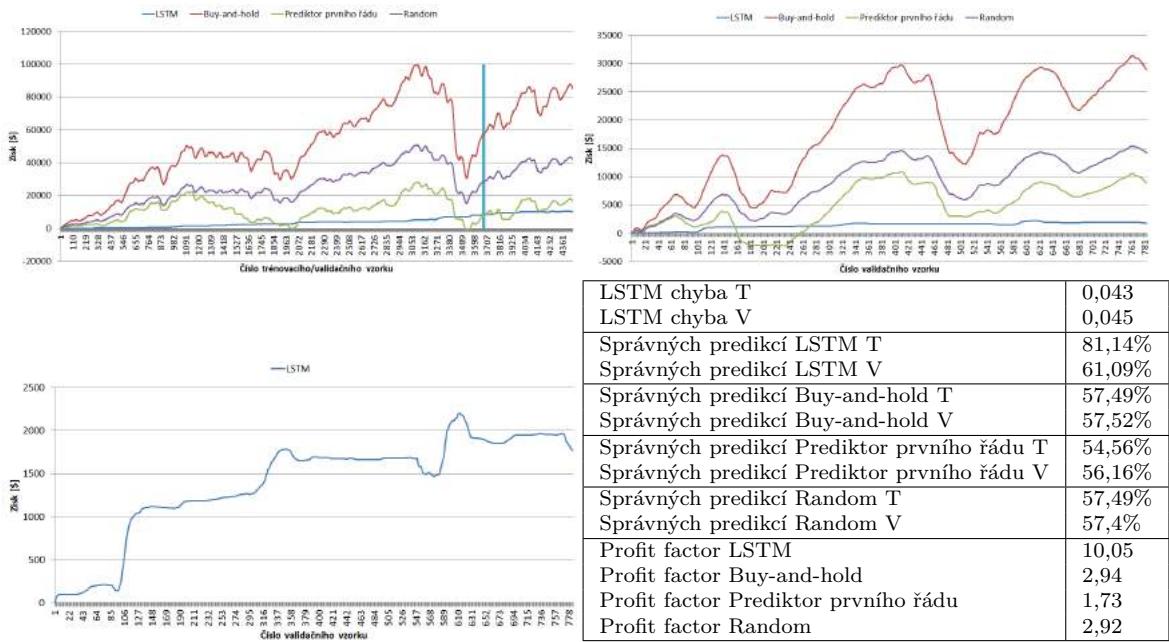
Stabilita strategie, stejně jako úspěšnost predikce, je mírně nižší. Profit factor byl sice zvýšen, nicméně křivka strategie je spíše skoková. Jelikož je dán důraz na stabilitu, je preferována spíše plynule rostoucí křivka. Další testy testovaly vliv počasí na jednotlivých kontinentech. Tyto výsledky jsou uvedeny na 9.2, 9.3, 9.4, 9.5 a 9.6.



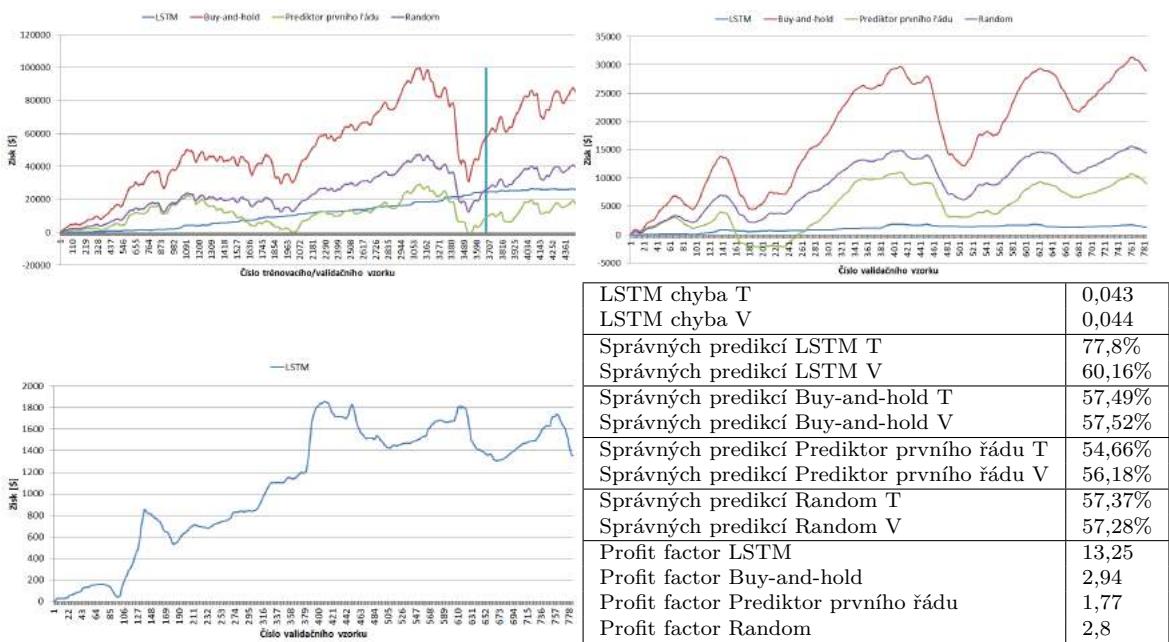
Obrázek 9.2: Přidání počasí v Evropě



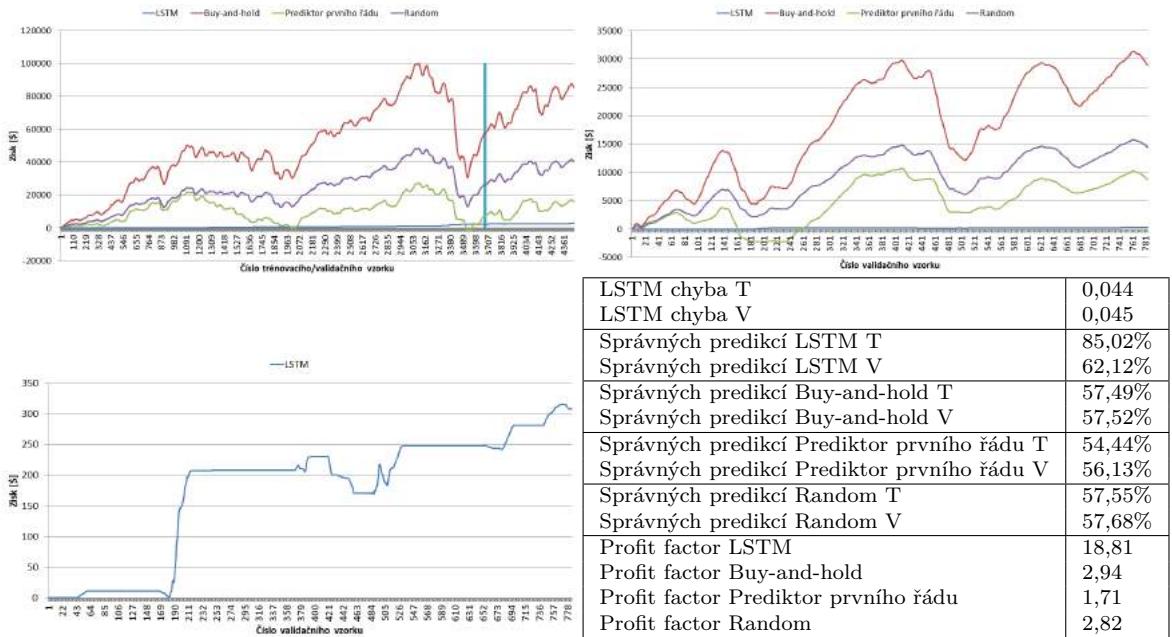
Obrázek 9.3: Přidání počasí v Severní Americe



Obrázek 9.4: Přidání počasí v Asii a Rusku



Obrázek 9.5: Přidání počasí v Jižní Americe

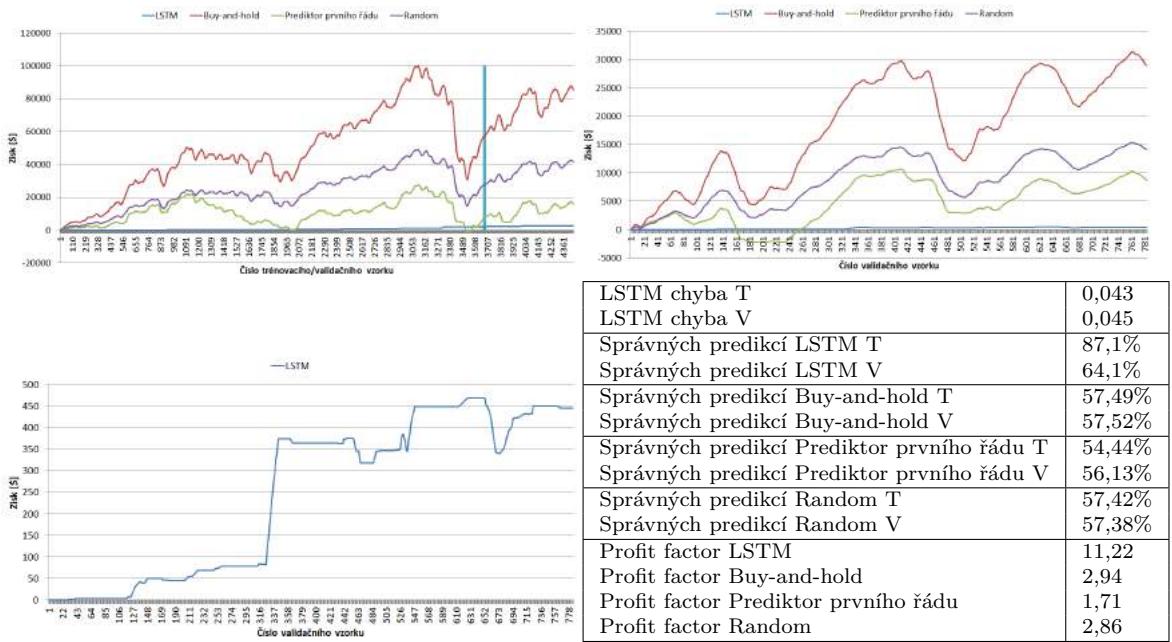


Obrázek 9.6: Přidání počasí v Africe

Z výše uvedených výsledků přineslo největší zlepšení počasí v Severní Americe. Tento výsledek je dobrý z hlediska úspěšnosti predikce a stability strategie na validačních datech. Chyba neuronové sítě je v těchto případech víceméně stejná. Profit factor je vyšší pouze při zapojení počasí v Africe, nicméně graf zisků není zdaleka tak plynule rostoucí. Výsledky se často výrazně liší ziskem na validačních datech. To je dáno pouze množstvím provedených obchodů¹. Nic ovšem nebrání tomu nakupovat akcii více a tím dostat zisk do libovolného rádu.

Kromě výše provedených testů byl zkoumán ještě vliv slunečního svitu. K tomu posloužily údaje *PSUN* a *TSUN* z dat o počasí. Výsledek tohoto testu (pro celý svět) je zobrazen na 9.7.

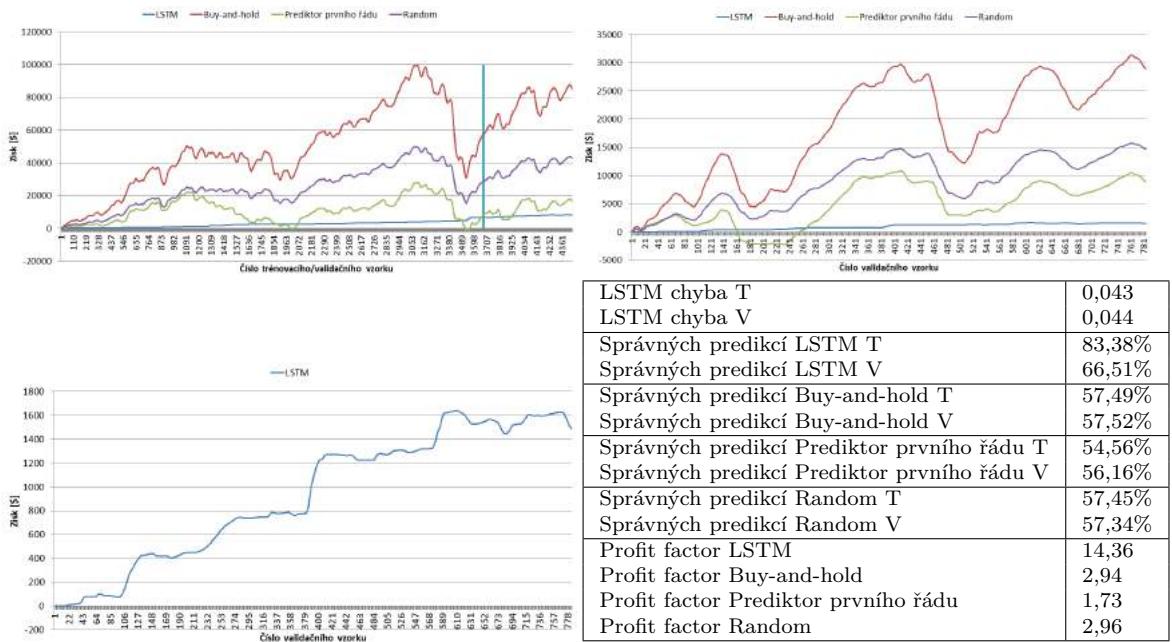
¹To je závislé na filtrování dle síly predikce.



Obrázek 9.7: Přidání slunečního svitu

Výsledek je mírně lepší než referenční výsledek (8.19), ale stále slabší než vliv počasí v Severní Americe. Rozdíl je hlavně ve stabilitě obchodní strategie. Dále proto bude uvažováno pouze počasí v Severní Americe.

Jelikož Wall Street se nachází v New Yorku a New York, stejně jako další významná města (Washington), leží na východním pobřeží USA, tak v dalším testu bylo počasí Severní Ameriky omezenou pouze na východní pobřeží USA. Výsledek tohoto testu je na 9.8.

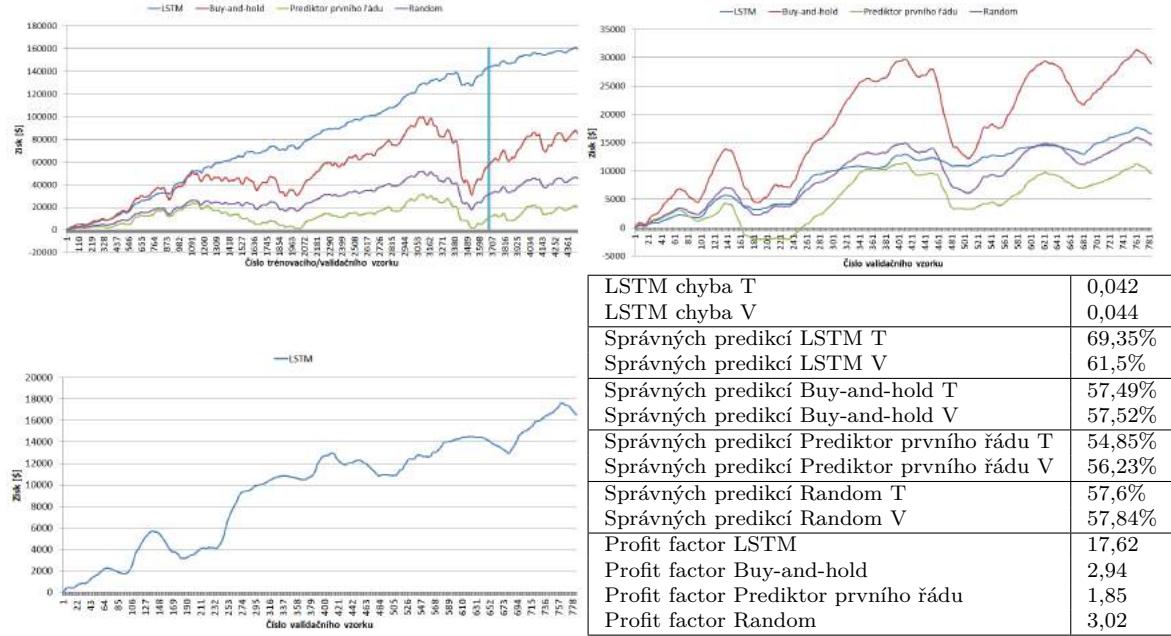


Obrázek 9.8: Přidání počasí na východním pobřeží USA

Výsledek tohoto testu opět vylepšil výsledky. Významně stoupala úspěšnost predikce a

obchodní strategie je na validačních datech velmi stabilní. Oproti ostatním strategiím je méně zisková, ale toto je opět díky menšímu množství obchodů. Profit factor poklesl pouze mírně.

Nakonec byla ještě otestována metoda PCA. Výsledek (bez redukce) je zobrazen na 9.9.



Obrázek 9.9: Přidání počasí na východním pobřeží USA s PCA

Výsledek s PCA trpí o hodně nižší úspěšností predikce, ačkoliv chyba neuronové sítě je prakticky stejná a profit factor dokonce vyšší. Přesto je strategie na validačních datech stále celkem stabilní. Ve srovnání s 9.8 je tento výsledek ale celkově horší, proto metoda PCA nebude používána. Ani test s PCA s redukcí (50%) nepřinesl žádné zlepšení, proto zde není uveden.

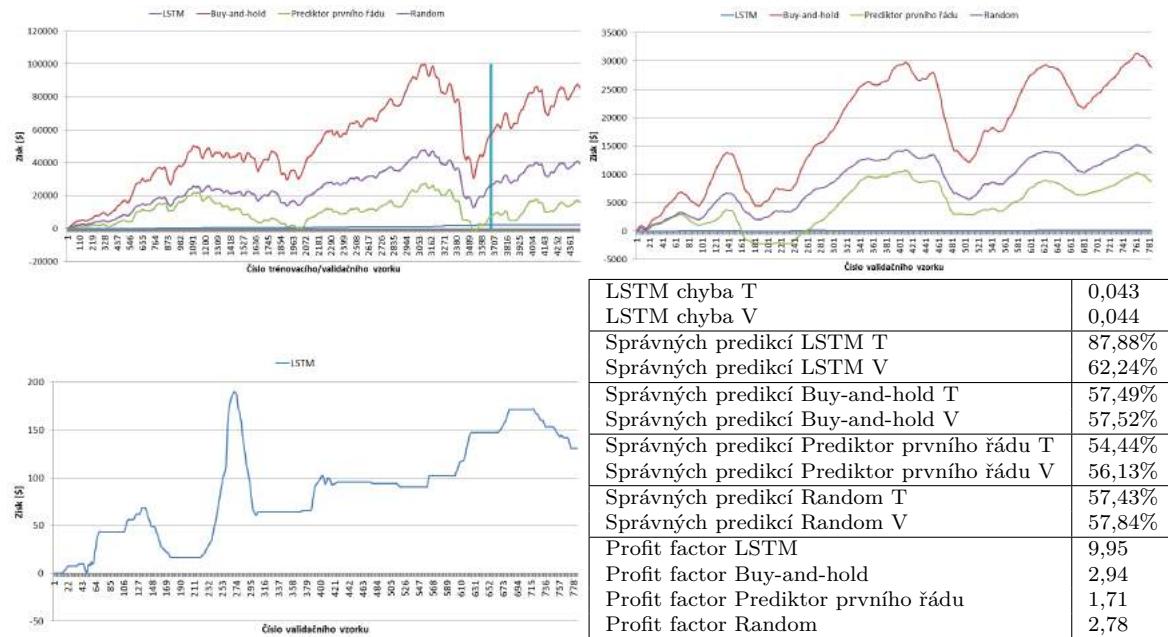
Úspěšnost obchodní strategie se tedy podařilo zvýšit pomocí informací o počasí na východním pobřeží USA. Nejlepší dosažený výsledek v této kapitole je tedy zobrazen na 9.8. Na závěr je ještě uvedena tabulka konfigurace (9.1).

Vstupní data	historické ceny akcií, počasí na východním pobřeží USA
Chybějící hodnoty	náhodné hodnoty
Vstup	pokles/růst za posledních 30 dní
Výstup	pokles/růst následujících 30 dní, zakódováno do 0-1
Normalizace	z-score, bez okénka
PCA	ne
Velikost skryté vrstvy	30
Počet iterací	40
Dropout	vstupní vrstva - 0,8, skrytá vrstva - 0,5
Learning rate	0,01
Aktualizace learning rate	ano
Gradient clip	5
Velikost dávky (batch)	10
Obchodní systém - vstup	při predikci „dostatečného“ růstu
Obchodní systém - výstup	po 30 dnech

Tabulka 9.1: Konfigurace po testu počasí

9.2 Forex

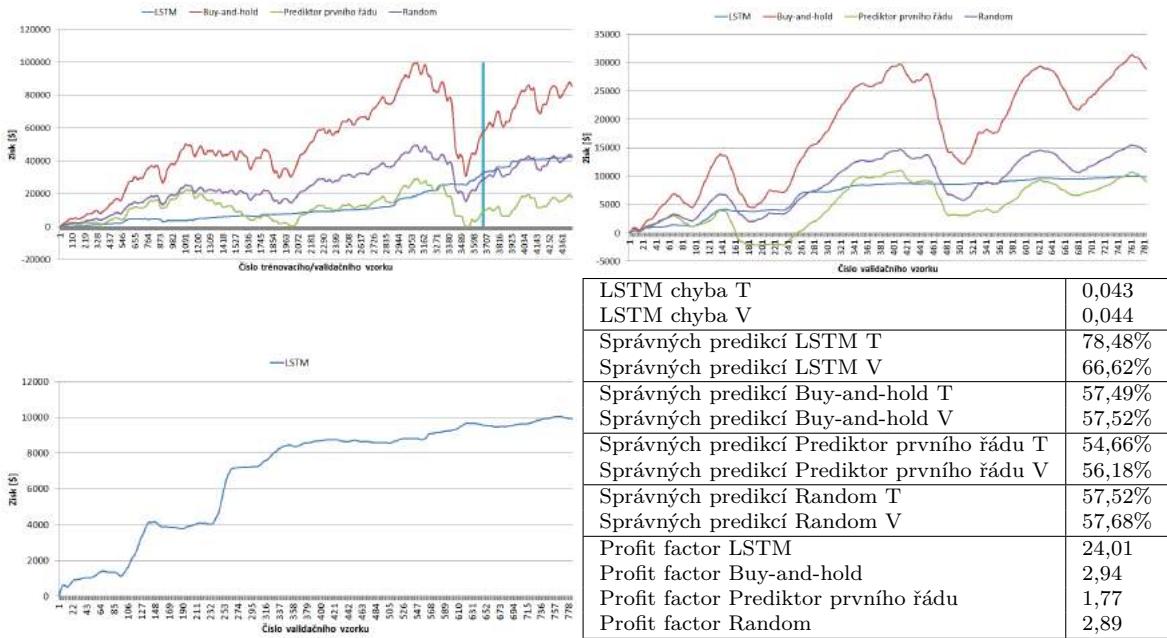
V dalším experimentu je zkoumán vliv Forexu na úspěšnost strategie. Jak bylo dříve zmíněno, jsou použity pouze měnové páry obsahující USD. Výsledek základního testu je zobrazen na 9.10.



Obrázek 9.10: Přidání forexových dat

Je naprosto zřejmé, že tato konfigurace vede na horší výsledky. V další fází tedy došlo k čištění těchto dat. Byly odstraněny měnové páry, kde chyběla více než polovina dat nebo které obsahovaly větší množství chybných dat². Díky tomuto čištění klesl počet datových řad Forexu cca na polovinu. Výsledek pro tento test je zobrazen na 9.11.

²Chybná data se projevovala nulovou změnou hodnoty.



Obrázek 9.11: Vliv Forexu po čištění dat

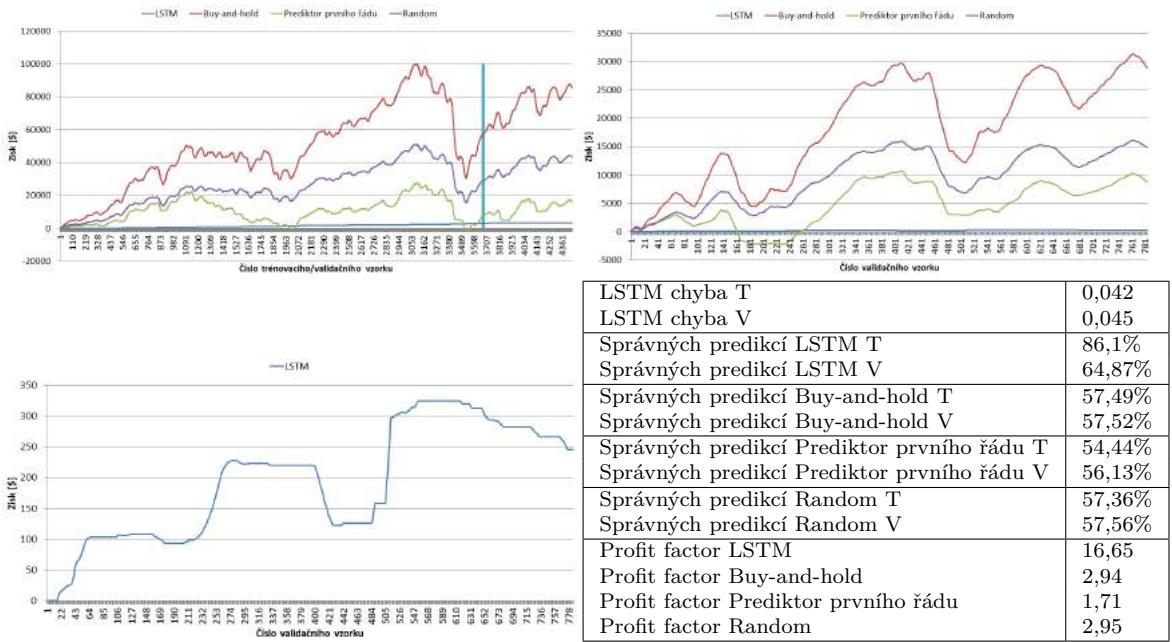
Po očištění dat už vykazuje strategie podstatně lepší výsledky. Úspěšnost predikce je prakticky stejná. Stabilita strategie je na validačních datech pouze mírně lepší. Avšak profit factor významně vzrostl. Forexová data tedy budou dále používána. Výsledná konfigurace se nachází v tabulce 9.2.

Vstupní data	
Chybějící hodnoty	historické ceny akcií, počasí na východním pobřeží USA, forexová data
Vstup	náhodné hodnoty
Výstup	pokles/růst za posledních 30 dní
Normalizace	pokles/růst následujících 30 dní, zakódováno do 0-1
PCA	z-score, bez okénka
Velikost skryté vrstvy	ne
Počet iterací	30
Dropout	40
Learning rate	vstupní vrstva - 0,8, skrytá vrstva - 0,5
Aktualizace learning rate	0,01
Gradient clip	ano
Velikost dávky (batch)	5
Obchodní systém - vstup	10
Obchodní systém - výstup	při predikci „dostatečného“ růstu po 30 dnech

Tabulka 9.2: Konfigurace po přidání Forexu

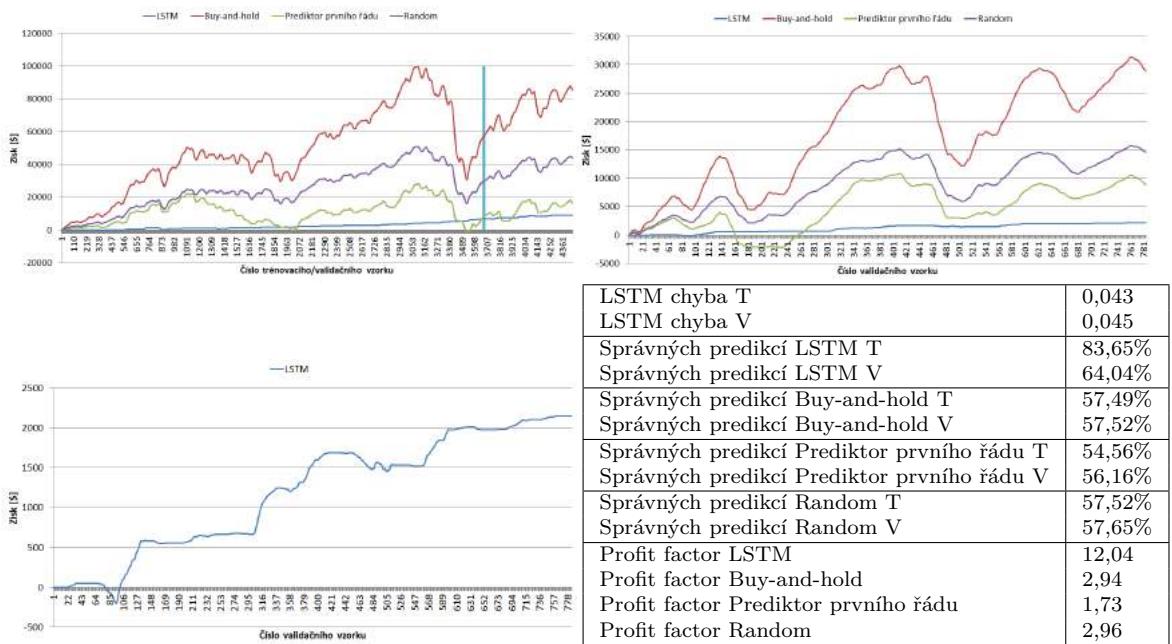
9.3 Google Trends

Testy v této kapitole zahrnují otestování přínosu dat pocházejících z Google Trends. Pokud připojíme data ke stávajícím zdrojům tak, jak byla v této práci definovaná, dostaneme výsledek zobrazený na 9.12.

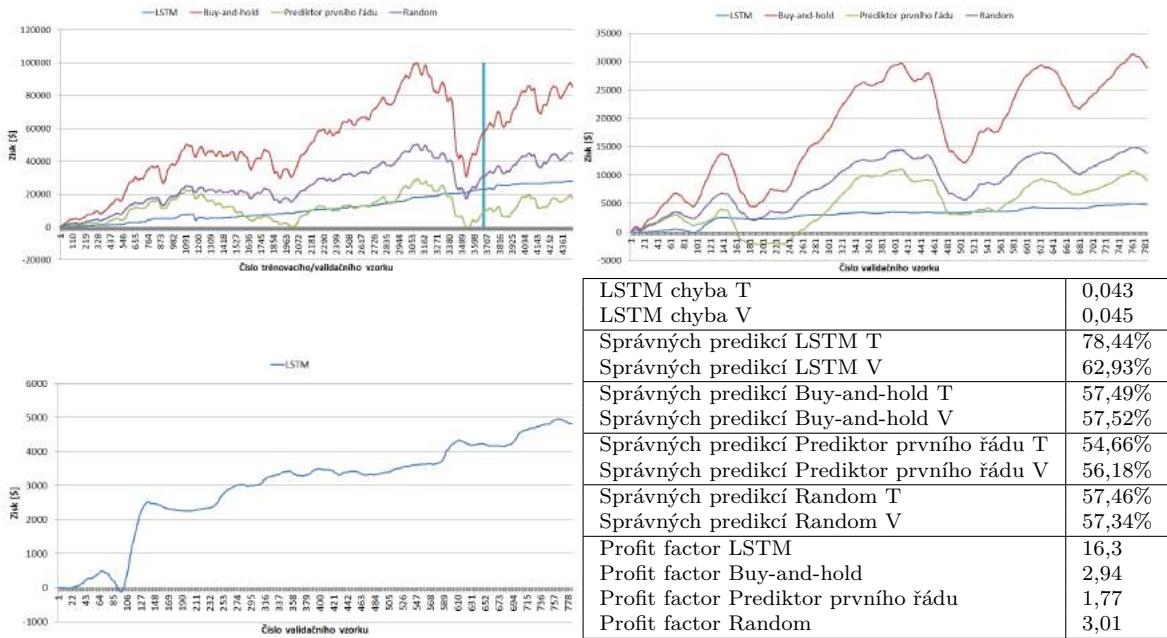


Obrázek 9.12: Přidání Google Trends

Z výše uvedených výsledků jasné plyne celkové zhoršení oproti 9.11. Jelikož se data z Google Trends skládají ze dvou typů dat (názvy firem a vybraná slova), tak dojde k testu jednotlivých typů dat. Tyto testy jsou zobrazeny na 9.13 a 9.14.



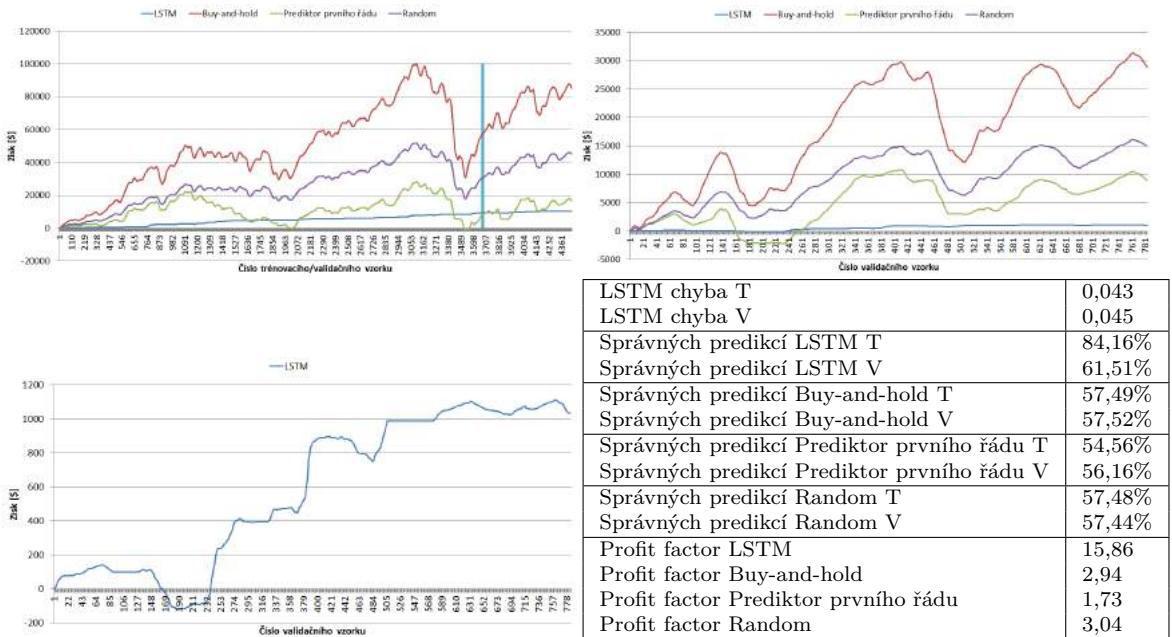
Obrázek 9.13: Přidání Google Trends omezeno pouze na názvy firem



Obrázek 9.14: Přidání Google Trends omezeno pouze na vybraná klíčová slova

Bohužel ani při oddělení jednotlivých typů dat nedošlo ke zlepšení výsledků. Velmi klesla úspěšnost predikce, ačkoliv strategie se na validačních datech jeví stále celkem stabilní. Protože všechny datové řady stažené z Google Trends obsahují stejné množství dat, nelze data dále čistit podle množství dostupných dat.

Jelikož v podkapitole 4.5 bylo zmíněno, že data z Google Trends jsou pouze měsíční a vždy dochází k rozkopírování hodnoty do celého měsíce, tak v posledním testu nebylo rozkopírování použito. Místo toho byla data doplněna náhodnými daty, stejně jako u ostatních datových řad. Výsledek pro tento test je zobrazen na 9.15.

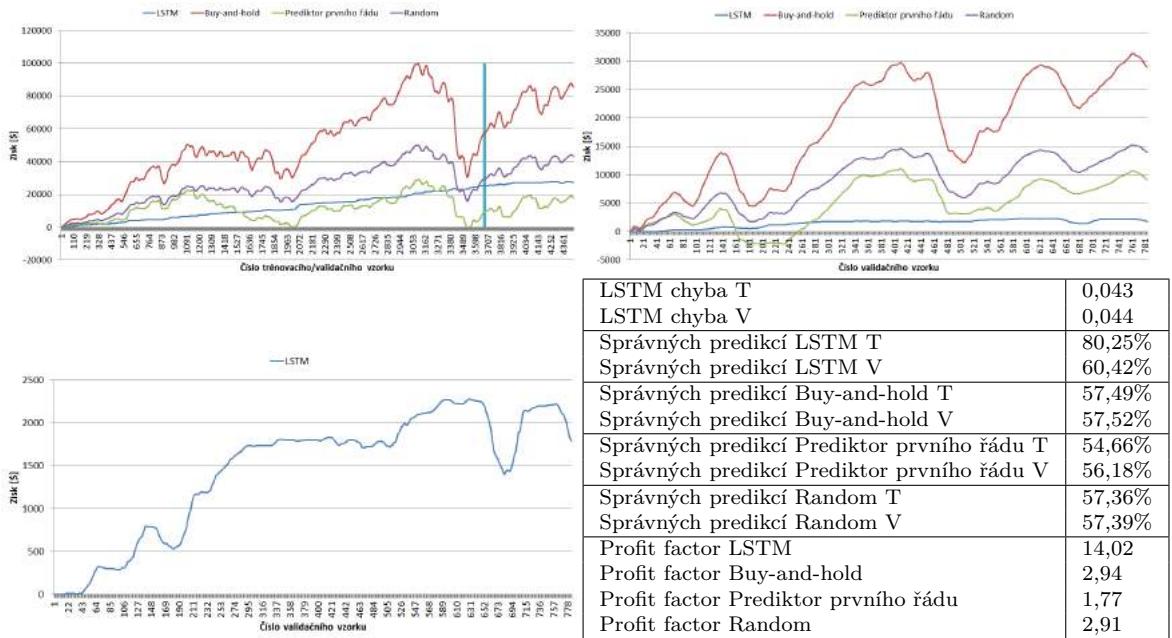


Obrázek 9.15: Google Trends při doplňování chybějících dat náhodnými daty

Avšak ani tento výsledek nepředčil dosud nejlepší dosažené výsledky. Díky datům z Google Trends se tedy nepodařilo strategii vylepšit a tato data nebudou dále používána. Konfigurace po těchto testech zůstava tak, jak je uvedena v tabulce 9.2.

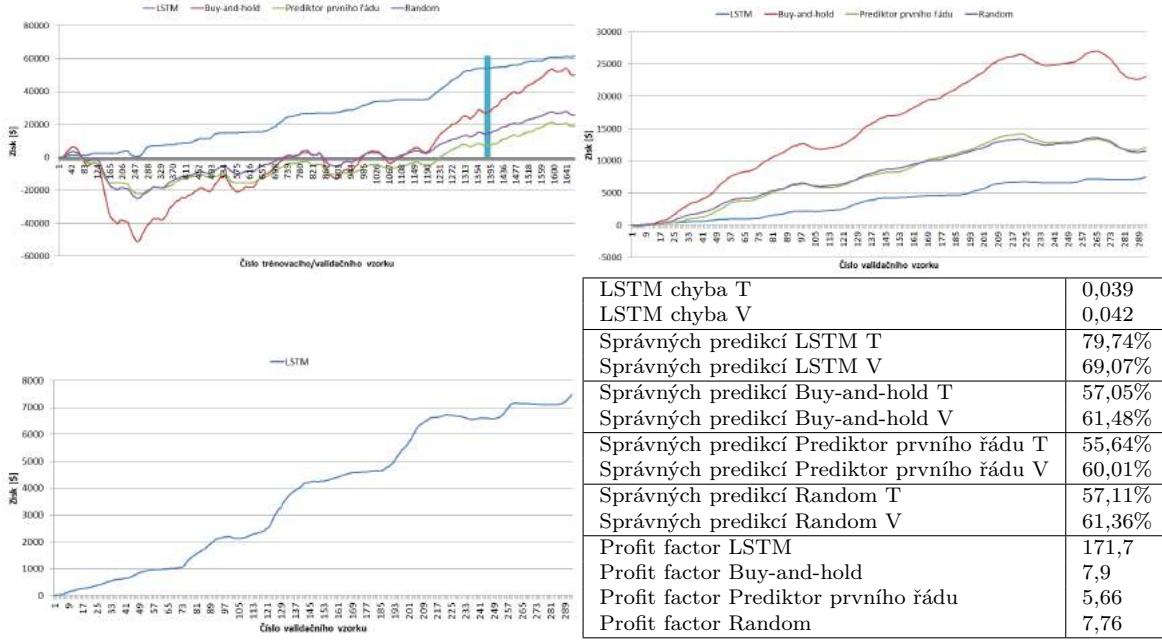
9.4 WikiTrends

Nyní jsou na řadě data z WikiTrends. Výsledek po přidání těchto dat je zobrazen na 9.16.

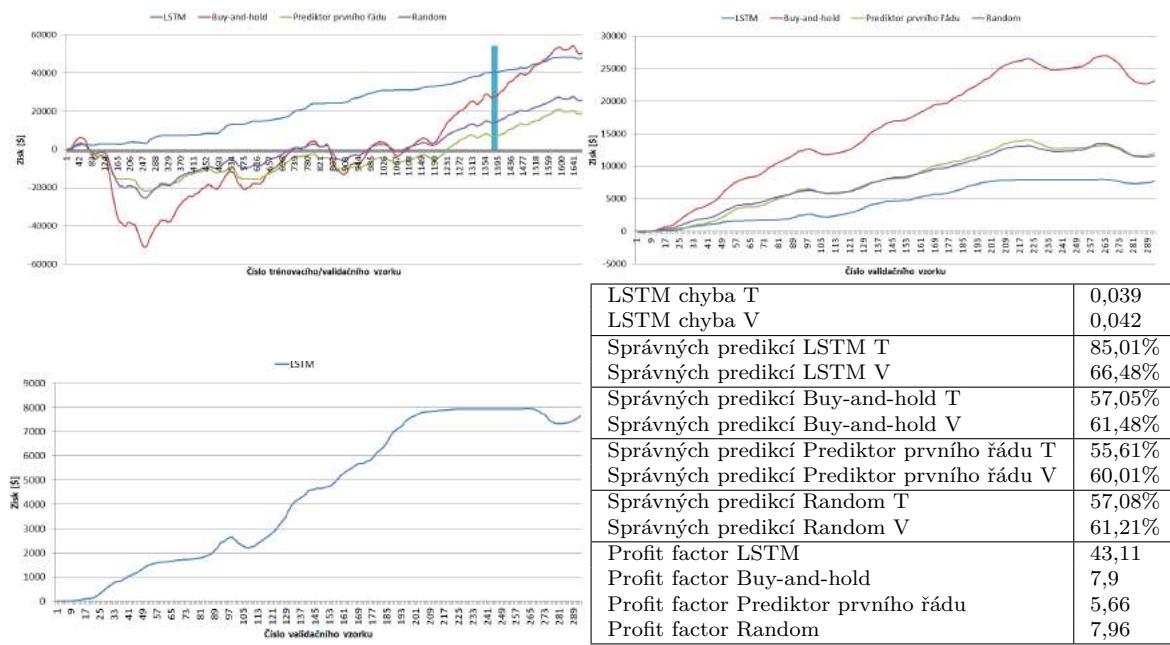


Obrázek 9.16: Přidání WikiTrends

Stejně jako v případě Google Trends, i teď došlo k významnému zhoršení výsledků. Data z WikiTrends jsou oproti Google Trends denní, ale problém je spíše jejich délka. Datové řady z WikiTrends začínají až od roku 2008. Přitom trénovací data při zvoleném rozdělení pokrývají interval od roku 1995 do roku asi 2009. To znamená, že WikiTrends se prakticky neúčastní procesu trénování. Pro otestování přínosu WikiTrends dat tedy změníme počátek uvažovaných dat na rok 2008. Výsledek pro aktuální konfiguraci (bez WikiTrends) je uveden na 9.17. Výsledek po přidání WikiTrends je pak zobrazen na 9.18.



Obrázek 9.17: Výsledek bez WikiTrends dat pro data od roku 2008



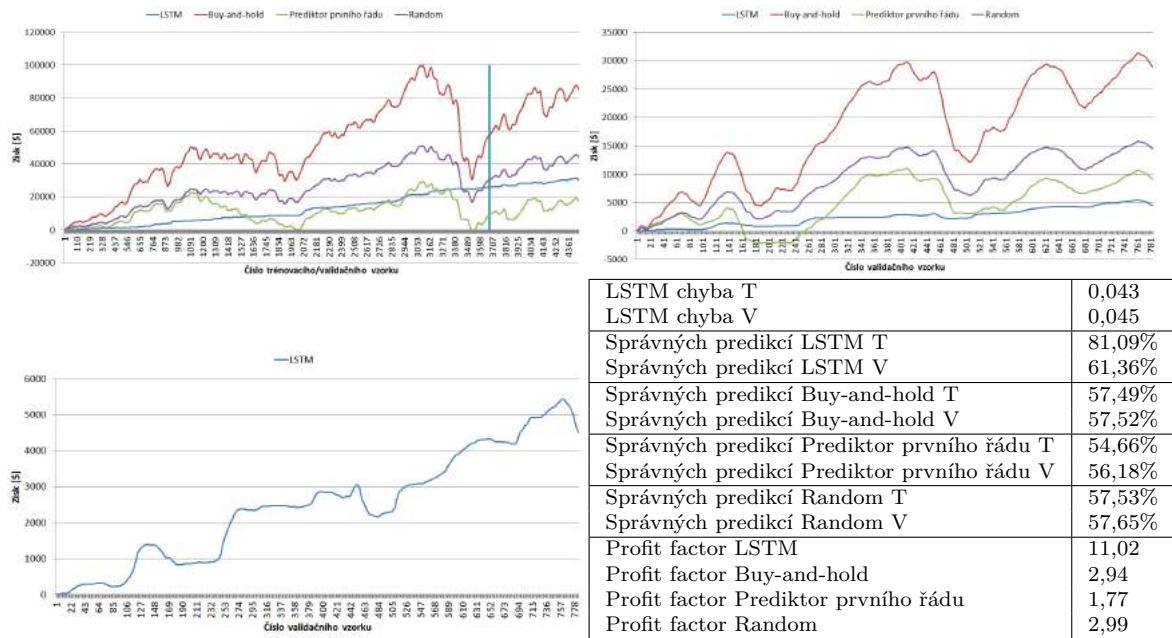
Obrázek 9.18: Výsledek s WikiTrends pro data od roku 2008

Obdobně jako v případě Google Trends, tak ani u WikiTrends se nepotvrdil pozitivní přínos těchto informací. Úspěšnost predikce i profit factor je nižší, stejně jako výkon strategie je méně stabilní. Chyba neuronové sítě zůstává klasicky velmi podobná. Dále byly testovány i jednotlivé typy dat (firmy, klíčová slova), ale výsledek byl opět stejný jak u Google Trends – ke zlepšení nedošlo.

Konfigurace se tedy opět nemění a data z WikiTrends nebudou při dalších testech uvažována.

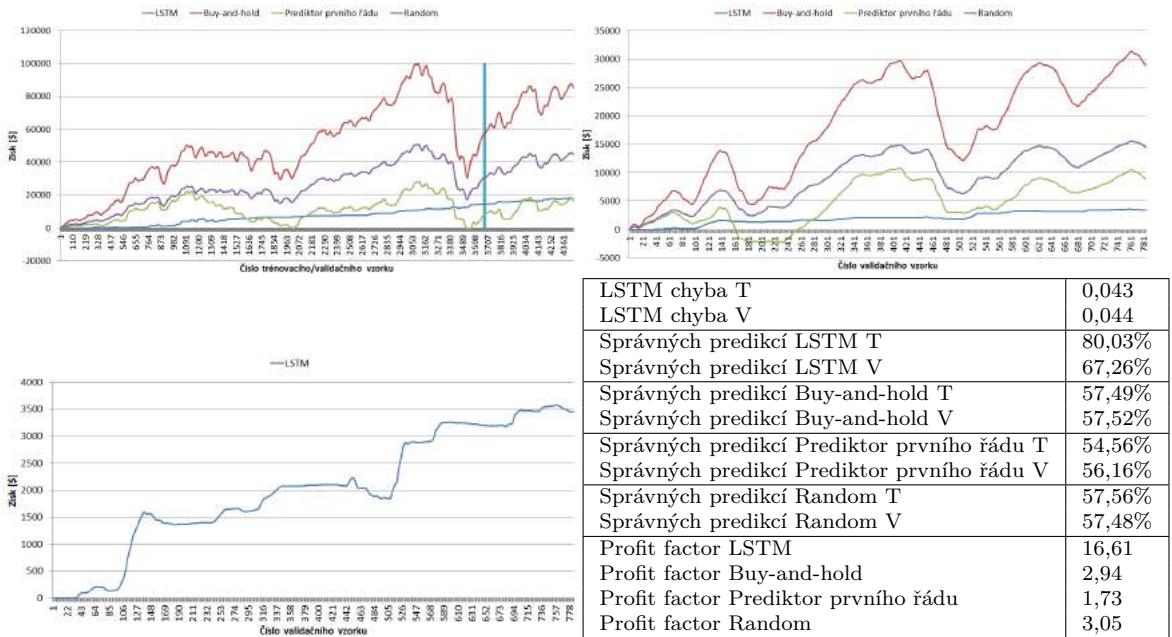
9.5 Futures

Předposledním testovaným zdrojem dat jsou futures data. Opět nejdříve proběhl základní test, jehož výsledek je zobrazen na 9.19.



Obrázek 9.19: Přidání futures dat

Celkový výsledek opět jednoznačně indikuje zhoršení. Data proto byla opět vyčištěna o datové řady, které obsahovaly větší množství chybějících dat. Tento výsledek je zobrazen na 9.20.



Obrázek 9.20: Přidání očištěných futures dat

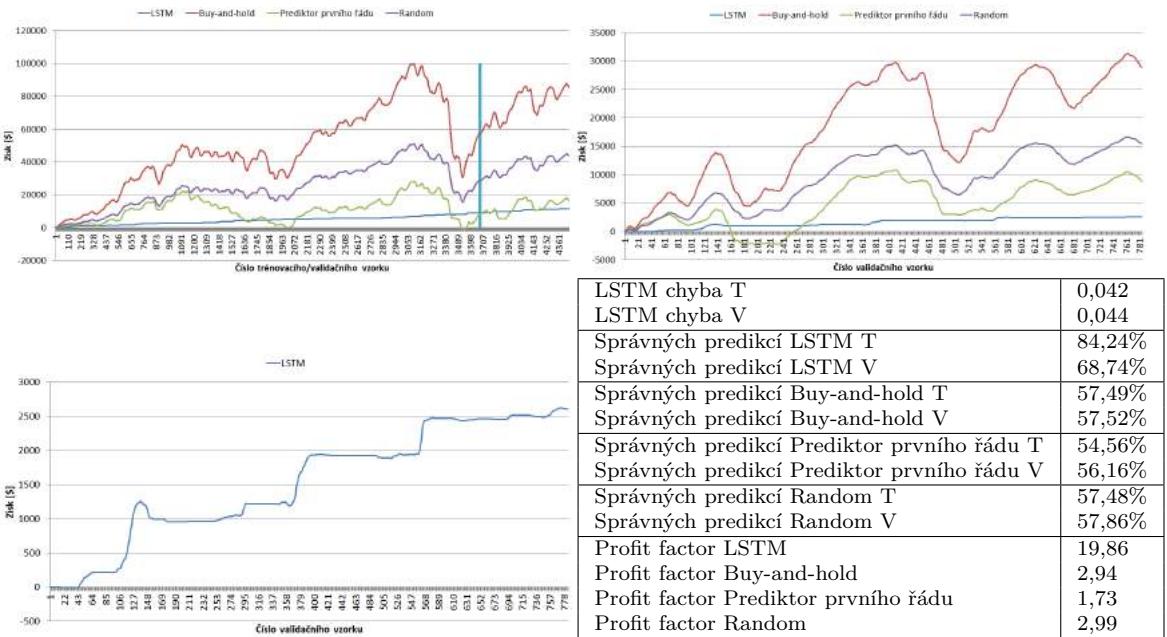
Výsledky byly mírně vylepšeny. Na chybě neuronové sítě jsou opět rozdíly minimální, ale vzrostla úspěšnost predikce. Výkon strategie je na validačních datech stále stabilní. Profit factor oproti předchozí konfiguraci poklesl. Dále se bude pracovat i s futures daty. Důvodem je zejména plynulejší růst zisku strategie. V předchozí konfiguraci byl růst strategie velmi nerovnoměrný. Tabulka 9.3 představuje novou konfiguraci.

Vstupní data	
Chybějící hodnoty	historické ceny akcií, počasí na východním pobřeží USA, forexová data, futures data
Vstup	náhodné hodnoty
Výstup	pokles/růst za posledních 30 dní
Normalizace	pokles/růst následujících 30 dní, zakódováno do 0-1
PCA	z-score, bez okénka
Velikost skryté vrstvy	ne
Počet iterací	30
Dropout	40
Learning rate	vstupní vrstva - 0,8, skrytá vrstva - 0,5
Aktualizace learning rate	0,01
Gradient clip	ano
Velikost dávky (batch)	5
Obchodní systém - vstup	10
Obchodní systém - výstup	při predikci „dostatečného“ růstu po 30 dnech

Tabulka 9.3: Konfigurace po přidání futures

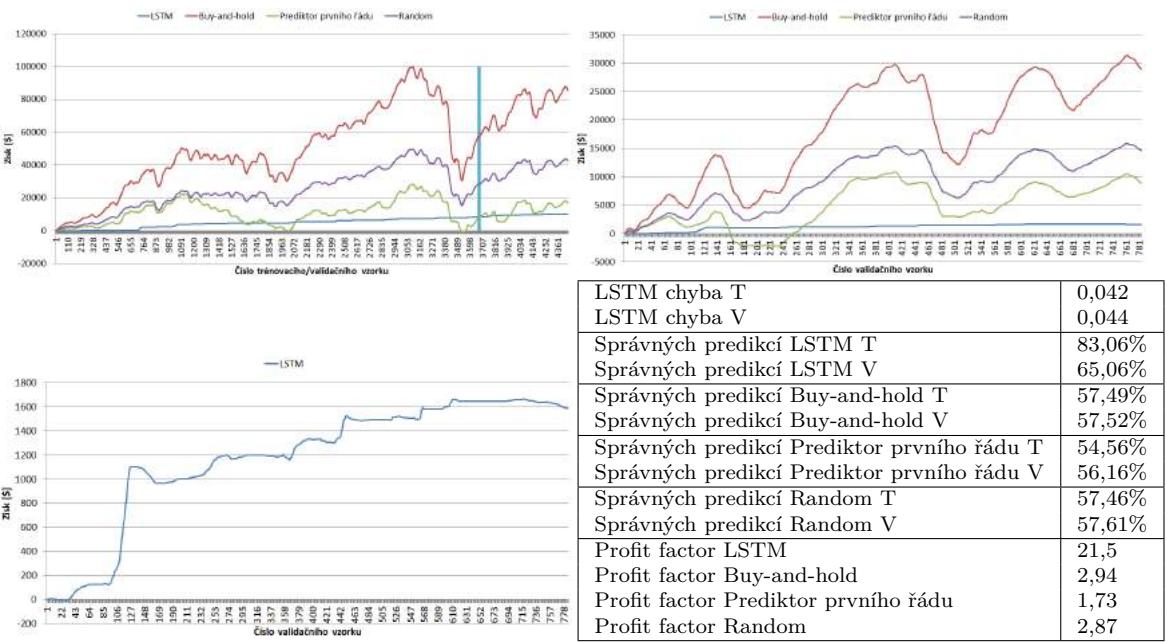
9.6 Fundamentals

Posledním zdrojem dat jsou fundamentální data. Tato data jsou na tom z hlediska kvality nejhůře. Většina datových řad obsahuje data s rozlišením jednoho roku. V této práci jsou ale používána data s denním rozlišením. Jak bylo popsáno v podkapitole 4.8, data s nižším časovým rozlišením jsou rozkopírována. Výsledek po přidání fundamentálních dat je zobrazen na 9.21.



Obrázek 9.21: Přidání fundamentálních dat

Výsledků ukazují (překvapivě) na zlepšení výkonnosti. Vzrostla úspěšnost predikce i profit factor a výsledky strategie jsou velmi stabilní. V dalším testu budou data očištěna o datové řady obsahující větší množství chybějících dat. Tento výsledek je zobrazen na 9.22.

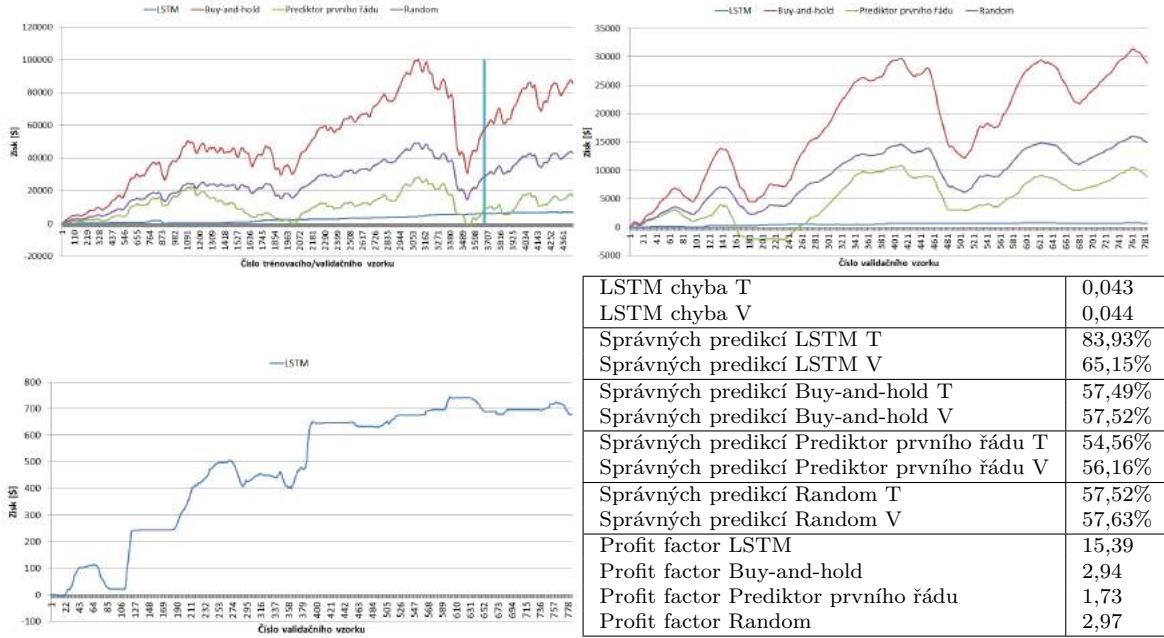


Obrázek 9.22: Přidání očištěných fundamentálních dat

Bohužel očištění dat vůbec nevedlo na zlepšení. Strategie je méně stabilní, ale hlavně velmi klesla úspěšnost predikce.

Podobně jak u Google Trends, i tady byl testován postup, kdy data nebyla pro svoji periodu rozkopírována, ale místo toho s nimi bylo zacházeno jak s běžnými chybějícími daty.

Výsledek je zobrazen na 9.23.



Obrázek 9.23: Přidání fundamentálních dat bez rozkopírování

Jak je vidět, ani tato úprava nezlepšila výsledky. Nejlepší výsledky tedy byly dosaženy s neočištěnými fundamentálními daty. Tato data tedy budou takto přidána do aktuální konfigurace, která je zobrazena v tabulce 9.4.

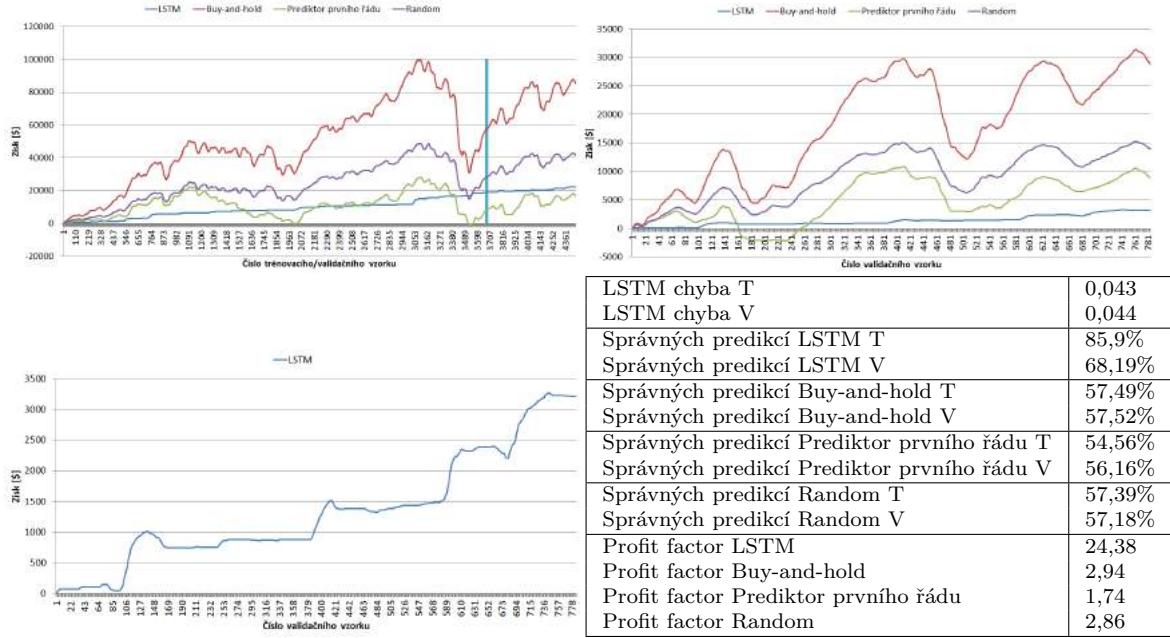
Vstupní data	historické ceny akcií, počasí na východním pobřeží USA, forexová data, futures data, fundamentální data
Chybějící hodnoty	náhodné hodnoty
Vstup	pokles/růst za posledních 30 dní
Výstup	pokles/růst následujících 30 dní, zakódováno do 0-1
Normalizace	z-score, bez okénka
PCA	ne
Velikost skryté vrstvy	30
Počet iterací	40
Dropout	vstupní vrstva - 0,8, skrytá vrstva - 0,5
Learning rate	0,01
Aktualizace learning rate	ano
Gradient clip	5
Velikost dávky (batch)	10
Obchodní systém – vstup	při predikci „dostatečného“ růstu
Obchodní systém – výstup	po 30 dnech

Tabulka 9.4: Konfigurace po přidání fundamentálních dat

9.7 Obchodní systém

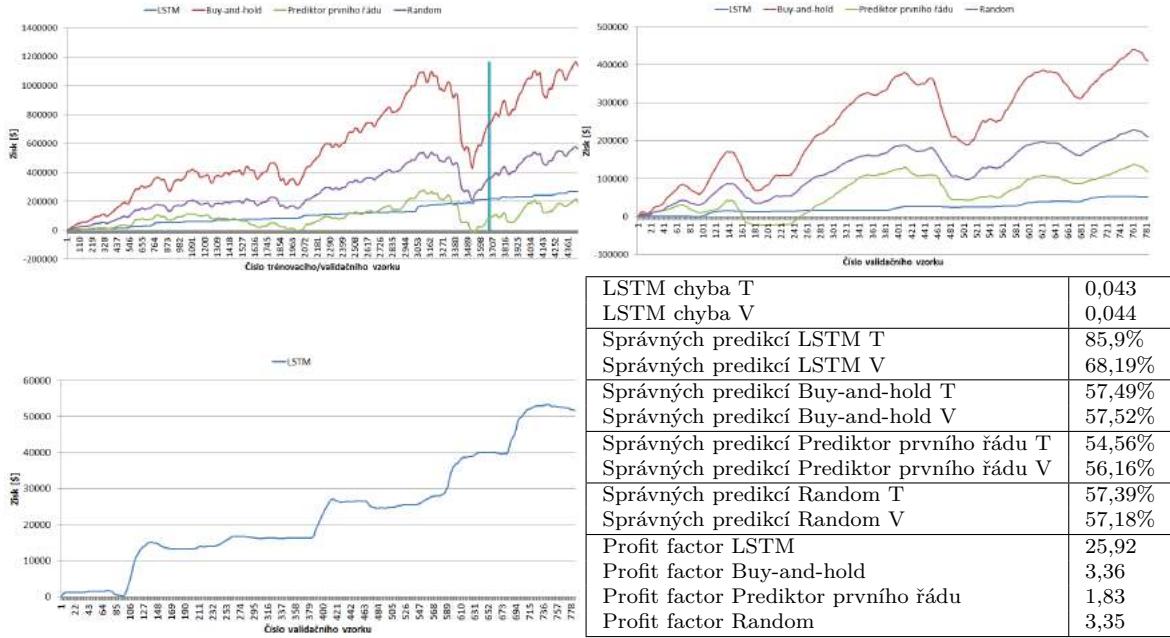
V této části budou testovány vybrané obchodní systémy. V aktuální konfiguraci dochází k predikci růstu/poklesu během příštích 30 dní. Při predikci „dostatečně velkého“ růstu dojde k nákupu 1 akcie. Pro srovnání bude sloužit výsledek 9.24 pocházející z aktuální konfigurace. K trénování a predikci došlo při těchto testech pouze jednou. Veškeré změny

budou totiž pouze v obchodním systému a proto bude stačit každý test provést pouze jednou.



Obrázek 9.24: Základní obchodní strategie

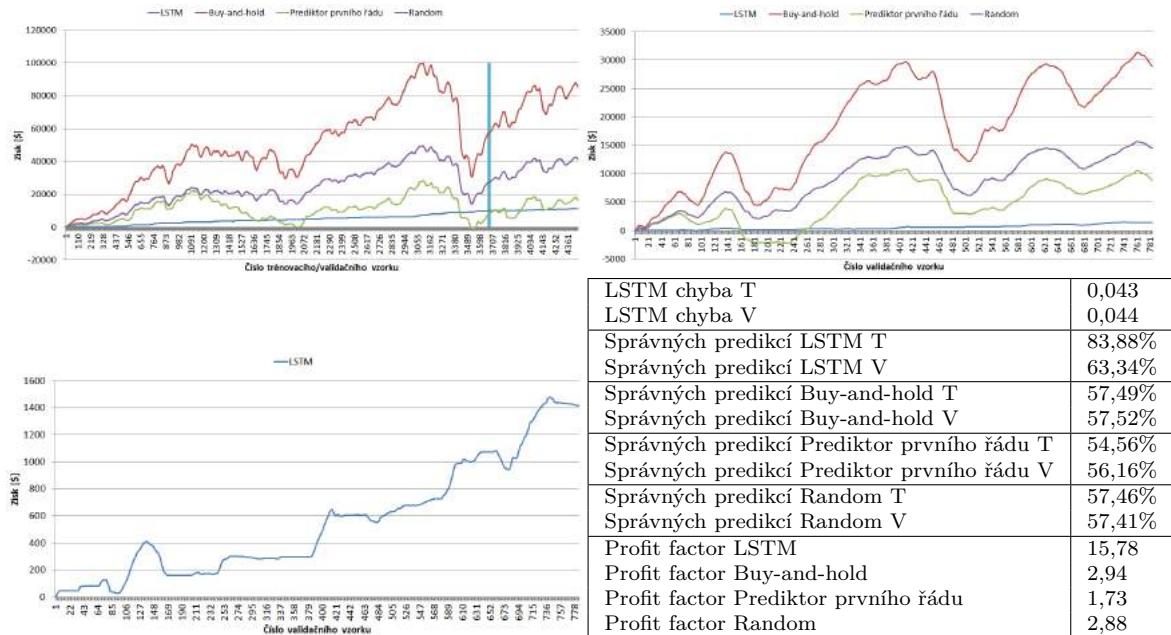
V prvním testu bude použito tzv. vyvažování akcií. To znamená, že pro každou firmu se určí průměrná velikost pohybu její akcie a podle toho se nakoupí počet akcií této firmy tak, aby každá akcie měla stejnou váhu. Tedy akcií s menším pohybem bude nakoupeno více než akcií s větším pohybem. Výsledek vyvažování akcií je zobrazen na 9.25.



Obrázek 9.25: Vyvažování akcií

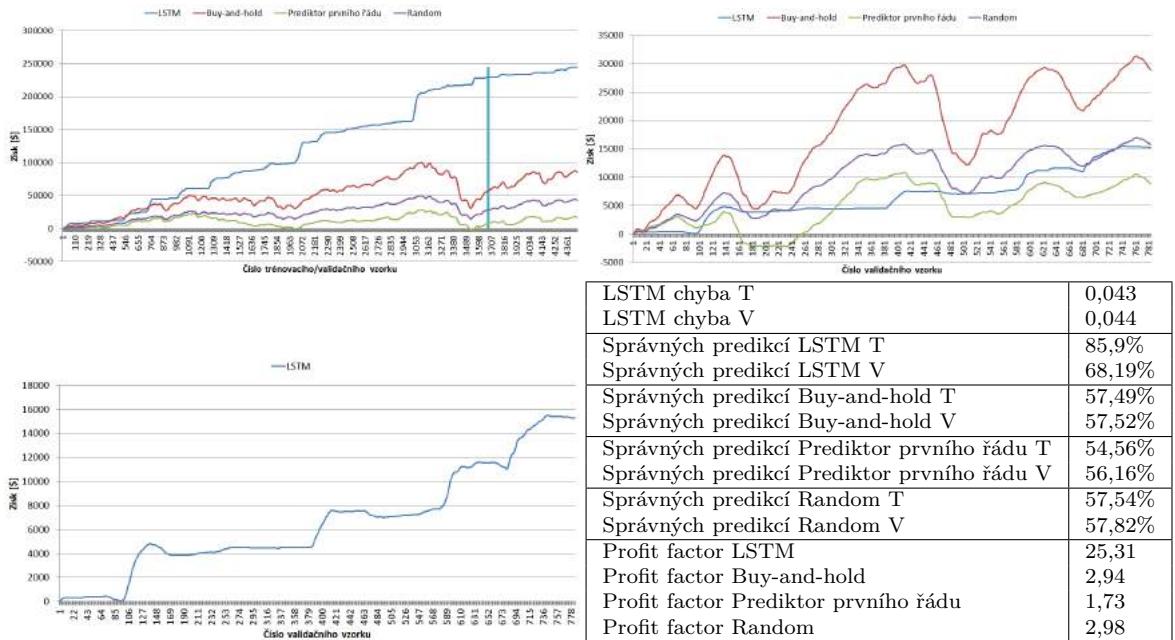
Počet správných predikcí se samozřejmě nezměnil. Došlo pouze ke změně křivky obchodní strategie. Ta je nepochybně o něco stabilnější. Kromě toho ale bylo dosaženo většího profitu. To je pouze důsledkem toho, že se často kupuje více než 1 akcie. Logicky je tato obchodní strategie mnohem náročnější na výši kapitálu, přitom dosažené zlepšení není zas tak výrazné.

Nyní bude navržen další přístup pro obchodování. Na každý den bude alokováno určité množství kapitálu. V tomto případě se lze rozhodovat různými způsoby. Lze nakupovat pouze nejlevnější akcie, lze nakupovat pouze akcie s predikcí nejvyššího růstu, nebo lze tyto přístupy kombinovat. Tento test bude zaměřen na výběr akcií podle nejvyšší predikce růstu. K nákupu bude vybráno 5 akcií s nejsilnější predikcí. Síla predikce ale opět musí být nad zvolenou hranicí. Výsledek testu tohoto obchodního systému je zobrazen na 9.26.



Obrázek 9.26: Koupě 5 akcií s nejsilnější predikcí

Při tomto přístupu došlo k výraznému snížení úspěšnosti predikce i profit factoru. Bude tedy otestován jiný přístup. Na každou akcií se vyhradí určitá částka a ta se za tuto akcií proinvestuje. Dražších akcií tedy bude nakoupeno méně než levnějších akcií. Částka k proinvestování byla zvolena jako cena nejdražší akcie. Výsledek je zobrazen na 9.27.



Obrázek 9.27: Nákup počtu akcií podle jejich ceny

Výsledky tohoto přístupu jsou velmi podobné s vyvažováním akcií. Úspěšnost predikce je opět nezměněná (musí být) a stabilita strategie je jen o trochu horší. Jedná vlastně o podobný přístup, pouze v tomto případě se počet akcií určuje podle ceny akcie a nikoliv podle velikosti průměrného pohybu.

Jelikož výsledky pro vyvažování akcií dopadly nejlépe, bude tento přístup uplatňován i dále. Aktuální konfigurace je uvedena v tabulce 9.5.

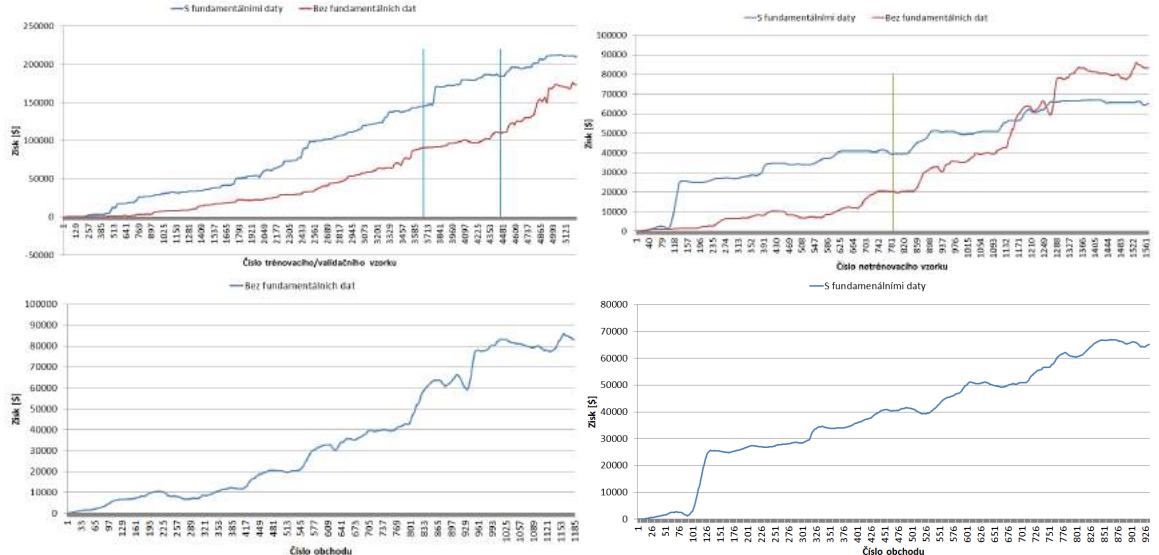
Vstupní data	
Chybějící hodnoty	historické ceny akcií, počasí na východním pobřeží USA, forexová data, futures data, fundamentální data
Vstup	náhodné hodnoty
Výstup	pokles/růst za posledních 30 dní
Normalizace	pokles/růst následujících 30 dní, zakódováno do 0-1
PCA	z-score, bez okénka
Velikost skryté vrstvy	ne
Počet iterací	30
Dropout	40
Learning rate	vstupní vrstva - 0,8, skrytá vrstva - 0,5
Aktualizace learning rate	0,01
Gradient clip	ano
Velikost dávky (batch)	5
Obchodní systém – vstup	10
Obchodní systém – výstup	při predikci „dostatečného“ růstu, množství akcie dle velikosti průměrného pohybu po 30 dnech

Tabulka 9.5: Konfigurace po změně obchodního systému

9.8 Vyhodnocení

Nyní dojde k vyhodnocení přínosu zapojení fundamentálních dat. Výsledky jsou shrnutý na 9.28. V prvním kroku byla pro každou strategii (s fundamentálními daty vs. bez fundamen-

tálních dat) nalezena nejvýnosnější konfigurace³. Právě proto jsou výsledky na validačních datech lepší než veškeré dosud zmiňované výsledky. Ty totiž vždy představovaly „průměrný“ výkon. Kromě toho jsou zde výsledky zobrazeny také na testovacích datech. V prvním grafu vertikální čáry oddělují postupně výkon strategie na trénovacích, validačních a nakonec testovacích datech. V druhém grafu jsou zobrazeny výsledky pouze na validačních a testovacích datech. Na těchto datech totiž neuronová síť nebyla učena. Avšak nejlepší konfigurace byla vybrána podle výkonu právě na validačních datech, kvůli dodržení *in-sample* a *out-of-sample* způsobu testování. Poslední dva grafy zobrazují výkon strategie s obchody na horizontální ose. Tyto grafy tedy vynechávají části na validačních a testovacích datech kde se zisk neměnil (v daný den nebyl proveden žádný obchod). Uvedené tabulky popisují výkon určité strategie postupně na trénovacích, validačních a nakonec validačních a testovacích datech dohromady.



LSTM chyba T	0,043	LSTM chyba T	0,042
LSTM chyba V	0,045	LSTM chyba V	0,044
Správných predíkcí LSTM T	79,6%	Správných predíkcí LSTM T	79,18%
Správných predíkcí LSTM V	69,86%	Správných predíkcí LSTM V	76,48%
Správných predíkcí LSTM	64,92%	Správných predíkcí LSTM	72,21%
Správných predíkcí Buy-and-hold T	57,49%	Správných predíkcí Buy-and-hold T	57,49%
Správných predíkcí Buy-and-hold V	57,52%	Správných predíkcí Buy-and-hold V	57,52%
Správných predíkcí Buy-and-hold	57,96%	Správných predíkcí Buy-and-hold	57,96%
Správných predíkcí Prediktor prvního řádu T	54,56%	Správných predíkcí Prediktor prvního řádu T	54,56%
Správných predíkcí Prediktor prvního řádu V	56,16%	Správných predíkcí Prediktor prvního řádu V	56,16%
Správných predíkcí Prediktor prvního řádu	56,74%	Správných predíkcí Prediktor prvního řádu	56,74%
Správných predíkcí Random T	57,62%	Správných predíkcí Random T	57,44%
Správných predíkcí Random V	57,69%	Správných predíkcí Random V	57,32%
Správných predíkcí Random	58,14%	Správných predíkcí Random	58,02%
Profit factor LSTM	3,46	Profit factor LSTM	5,26

Obrázek 9.28: Porovnání strategie bez fundamentálních dat a s fundamentálními daty

Už během testování byla vidět větší síla vyvíjené strategie oproti strategiím referenčním. Právě proto nejsou v grafech vyhodnocení zahrnutý. Uvedené výsledky vykazují zlepšení díky použití fundamentálních dat. Přesto je chyba neuronové sítě v obou případech velmi podobná. Nicméně tato chyba se příliš neměnila ani během testování. Úspěšnost predikce byla podstatně zvýšena. Co se stability obchodní strategie týče, ta byla díky fundamentál-

³Jedná se hlavně o dobré nastavení vah neuronové sítě při její inicializaci

ním datům velmi zvětšena. Strategie vykazuje rychlejší a pomalejší růsty, a zároveň pouze drobné propady. Obchodní strategie bez použití fundamentálních dat vykazuje také neutálý růst, její křivka je ale zubatější – je tedy méně stabilní. Profit factor se pohybuje oproti předchozím testům v jiných mezích, přesto taktéž vykazuje významné zlepšení. Za zmínku stojí také to, že nedošlo k přetrenování neuronové sítě. Výkon obchodní strategie je zhruba stejný na všech typech dat. Tento fakt je pro vývoj obchodní strategie velmi důležitý. Konečná konfigurace je uvedena v tabulce 9.6.

Vstupní data	historické ceny akcií, počasí na východním pobřeží USA, forexová data, futures data, fundamentální data
Chybějící hodnoty	náhodné hodnoty
Vstup	pokles/růst za posledních 30 dní
Výstup	pokles/růst následujících 30 dní, zakódováno do 0-1
Normalizace	z-score, bez okénka
PCA	ne
Velikost skryté vrstvy	30
Počet iterací	100
Dropout	vstupní vrstva – 0,8, skrytá vrstva – 0,5
Learning rate	0,01
Aktualizace learning rate	ano
Gradient clip	5
Velikost dávky (batch)	10
Obchodní systém – vstup	při predikci „dostatečného“ růstu, množství akcie dle velikosti průměrného pohybu po 30 dnech
Obchodní systém – výstup	

Tabulka 9.6: Výsledná konfigurace

Kapitola 10

Živé obchodování

Pro živé obchodování vyvájené strategie je potřeba najít brokeru, zajistit automatickou exekuci obchodních příkazů a získávat živá data. Kromě toho je potřeba spočítat veškeré náklady, požadavky na účet, risk apod.

Jako brokeru lze zvolit světově renomovanou společnost Interactive Brokers¹. Tato společnost nabízí obchodování s nejrůznějšími produkty, mj. akcemi. Zařazení dalších produktů do obchodovacího portfolia tedy není problém.

10.1 Technické provedení

Interactive Brokers nabízí API pro exekuci obchodních příkazů, získávání dat apod. Toto API je přístupné hned v několika programovacích jazycích – Java, C#, C++, ActiveX (Excel, Matlab) atd. Pro použití je potřeba stáhnout potřebné zdrojové kódy². Kromě zdrojových kódů tento balík obsahuje i ukázkové aplikace. Manuál pro C# API je dostupný na <https://individuals.interactivebrokers.com/en/software/csharp/c.htm>. Tímto je tedy zajištěno spojení obchodní strategie s brokerem. Poté zbývá upravit vyvájenou aplikaci pro použití s tímto API. Jelikož je zajištění dat, predikce LSTM neuronovou sítí a simulace obchodního systému důsledně oddělena, není třeba provádět velké úpravy.

10.2 Finanční záležitosti

Pro založení účtu u Interactive Brokers je potřeba na účet vložit minimálně \$10000. Během obchodování se tato částka může dostat i pod \$10000, ale je dobré mít na účtu stále určitou rezervu. Při obchodování akcií se potom účtuje poplatek 0,1% z hodnoty obchodu. Ovšem minimální hodnota poplatku je \$4 a maximální je \$29³. Přitom není třeba kupovat jednotlivé akcie zvlášť, ale lze nakoupit tzv. koš akcií⁴. Výběr peněz z IB účtu lze provést jednou měsíčně zadarmo. Při dalších výběrech je účtován poplatek 300 Kč. V České Republice ještě dochází ke zdanění příjmů, jak bylo uvedeno v části 2.3.3. Při krátkodobějších obchodech tedy dojde k zaplacení 15% daně z příjmu u fyzických osob nebo 19% u právnických osob.

Pro spočítání potřebného kapitálu uvažujme rok 2016, kdy je průměrná hodnota akcií vybraných 100 akcií nejvyšší. V této době je průměrná hodnota akcie cca \$50. Nebudeme

¹<https://www.interactivebrokers.com/en/home.php>

²<http://interactivebrokers.github.io/>

³<https://www.interactivebrokers.com/en/index.php?f=commission&p=stocks1>

⁴<https://www.interactivebrokers.com/en/?f=%2Fen%2Ftrading%2Forders%2Fbasket.php>

nyní uvažovat zmiňované vyvažování akcií, ale techniku, která vedla k podobným výsledkům – alokace fixní částky pro každou různou akci (tedy firmu). Jelikož cena nejvyšší akcie je cca \$150, přepokládejme že tato částka bude použita na každou různou akci. Ve zvolené obchodní strategii počet různých akcií vhodných k nákupu většinou nepřesáhne 20 ze 100. Při koupi akcií od 20 firem tedy dojde k nákupu akcií v celkové hodnotě $20 \cdot \$150 = \3000 . Celkem s poplatky činí náklady \$3003. Při spekulaci na vzrůst ceny n dní dopředu lze buď akcie nakoupit a po n dnech je prodat a poté proces opakovat, nebo lze každý den nakoupit koš akcií a po n dnech daný koš akcií zase prodat. V druhém případě se tedy každý den drží n akcií (pří nákupu o velikosti 1 akcie). Tento způsob byl vybrán při testování, neboť dochází k více obchodům a lépe se tedy posuzuje úspěšnost strategie. Při druhém způsobu a spekulaci na růst 30 dní dopředu je proto třeba disponovat minimálně $\$3003 \cdot 30 = \90090 . Při živém obchodování je samozřejmě potřeba mít k dispozici vyšší částky pro úhradu ztrát.

Kapitola 11

Závěr

Tato práce měla za úkol vytvořit obchodní systém obchodující akcie s využitím fundamentálních informací. Rozhodování bylo prováděno pomocí neuronové sítě. Jako vstupy tedy nesloužily pouze historické ceny (technická analýza), ale i fundamentální data (fundamentální analýza). Byly zkoumány různé zdroje fundamentálních dat. Výkon strategie byl úspěšně zvýšen pomocí počasí na východním pobřeží USA. To lze přisuzovat tomu, že toto pobřeží je poměrně hustě obydlené. Dále je to sídlo mnoha firem, ale hlavně se zde nachází proslulá Wall Street. Následoval test forexových dat a i ty se podařilo v obchodní strategii využít. Naproti tomu při použití dat z Google Trends a WikiTrends nedošlo ke zlepšení výkonnosti. Hlavní příčinou bylo určitě omezené množství dat a jejich časové rozlišení. Navíc data z WikiTrends byla dostupná až od roku 2008. Test futures dat opět přinesl zlepšení. Tato data postihují různá odvětví (energie, plodiny apod.) a proto slouží jako dobrý monitor globální ekonomiky. Posledními daty byla fundamentální data. V tomto případě byl očekáván negativní přínos, jelikož data měla velmi nízké časové rozlišení. Přesto v tomto případě došlo ke zlepšení výsledků.

Při celkovém vyhodnocení byl kladně vyhodnocen přínos fundamentální analýzy. Referenční obchodní strategie se pohybovaly v úspěšnosti predikce okolo 57%, vyvýjená obchodní strategie bez fundamentálních dat dosáhla úspěšnosti asi 64% a stejná obchodní strategie s fundamentálními daty nakonec dosáhla úspěšnosti až 72%. Ačkoliv jsou tato procenta mírně nepřesná¹, přesto lze pozorovat značné zlepšení díky fundamentální analýze. Důvodem je také to, že akcie právě fundamentálními daty bývají ovlivňovány častěji než třeba takové akciové indexy.

Kromě fundamentální analýzy byly v této práci testovány různé techniky (hlavně co se týče neuronových sítí) pro zlepšení predikce². Došlo například k implementaci dropoutu u rekurentní LSTM neuronové sítě. Tato technika významně snížila přetrénování neuronové sítě. Dále byla testována metoda PCA pro redukci dimenze, avšak její použití nepřineslo pozitivní výsledky. Kromě toho byly testovány dvě metody dlouhodobé predikce. V tomto případě se osvědčila právě metoda přímé dlouhodobé predikce. Dále byly testovány různé transformace dat pro lepší učení neuronové sítě. Nakonec byly ještě testovány různé metody normalizace. Kvůli charakteristice dat byl očekáván přínos okénkové normalizace, nicméně tato metoda vedla naopak ke zhoršení výsledků.

Fundamentální analýza tedy vedla ke zlepšení obchodního systému. Podrobné výsledky jsou uvedeny v podkapitole 9.8.

¹Úspěšnost velmi také závisí na inicializaci vašich neuronových sítí.

²Ne všechny tyto závěry jsou z uvedených testů tak zřejmé, nicméně byly neustále potvrzovány v průběhu vývoje.

11.1 Možnosti pokračování

Implementovaná strategie může být dále vylepšována. V této části budou zmíněny některé nápadы, které by mohly vést k lepším výsledkům.

- Může být použito více zdrojů fundamentálních dat. Kromě dat kvantitativních lze přidat i data získaná z různých textů a článků (po náležité transformaci dat). Hlavně by mohla být přidána data ze SEC³. Avšak tato data jsou velmi nestrukturována a jejich získání a očištění bude zřejmě velmi náročné. Alternativou je samozřejmě napojení na placené zdroje dat.
- Kromě zvýšení zdrojů fundamentálních dat je také důležité zajistit dostatečnou kvalitu dat a jejich délku.
- Je potřeba vylepšit inicializaci vah neuronové sítě, aby byly výsledky více konzistentní a vyhodnocení pak jednodušší.
- Vstupní data by mohla být ještě předzpracována (například vyfiltrována a vyhlazena), rozšířena o indikátory apod.
- Lze aplikovat také technickou analýzu pro zvýšení přesnosti predikce.
- Bylo by dobré stávající systém rozšířit o možnosti obchodovat i jiné produkty kromě akcií. Pak by mohl být zjištěn přínos fundamentálních informací i třeba pro akciové indexy.
- Učení neuronové sítě by šlo vylepšit průzkumem různých technik a jejich aplikací. To by mohlo vést i na úplnou změnu architektury neuronové sítě.
- Při živém nasazení implementovaného systému by bylo bezpodmínečně nutné řídit investiční riziko. V aktuální implementaci nic takového není.
- Aby výsledky odpovídaly realitě co nejvíce, je potřeba předejít *survivorship bias* (viz podkapitola 4.1). Pro tuto práci bohužel nebyla nalezena data, aby bylo možné této situaci předejít. Tato data nebývají veřejně dostupná a nachází se často právě v placených databázích.
- Řízení obchodu by mohlo být lépe řešeno. Nyní je délka obchodu napevno daná podle délky predikce. V případě ztráty by mohl být obchod dříve ukončen, aby ztráta nebyla dále zvětšována.

³<https://www.sec.gov/>

Literatura

- [1] ALEXANDRE, L.; MARQUES, J.: Error Entropy Minimization for LSTM Training. [Online], [cit. 2016-05-15].
URL <http://www.di.ubi.pt/~lfbaa/pubs/Alexandre-MarquesSa.pdf>
- [2] ATIYA, A.: Bankruptcy Prediction for Credit Risk Using Neural Networks: A Survey and New Results. *IEEE transactions on neural networks*, ročník 12, č. 4, 2001, ISSN 1045-9227.
- [3] Bank for International Settlements: Triennial Central Bank Survey of foreign exchange turnover in April 2013. 2013, [Online], [cit. 2015-12-17].
URL <http://www.bis.org/press/p130905.htm>
- [4] Board of Governors of the Federal Reserve System: What is the purpose of the Federal Reserve System? 2014, [Online], [cit. 2015-12-17].
URL http://www.federalreserve.gov/faqs/about_12594.htm
- [5] BROGAARD, J. A.: High Frequency Trading and its Impact on Market Quality. 2010, [Online], [cit. 2015-12-17].
URL <http://www.clasesdebolsa.com/archivos/HTF.pdf>
- [6] CME Group: S&P 500 Futures Contract Specs. 2015, [Online], [cit. 2015-12-17].
URL http://www.cmegroup.com/trading/equity-index/us-index/sandp-500_contract_specifications.html
- [7] Computer Science Department, Stanford University: Convolutional Neural Network. [Online], [cit. 2015-12-17].
URL <http://ufldl.stanford.edu/tutorial/supervised/ConvolutionalNeuralNetwork>
- [8] CONT, R.: Statistical Modeling of High-Frequency Financial Data. *Signal Processing Magazine*, ročník 28, č. 5, 2011, ISSN 1053-5888.
- [9] Dow Jones: Dow Jones History. [Online], [cit. 2015-12-17].
URL <http://solutions.dowjones.com/history.asp>
- [10] DUHOUX, M.; SUYKENS, J.; MOOR, B. D.; aj.: Improved long-term temperature prediction by chaining of neural networks. [Online], [cit. 2016-03-14].
URL <ftp://ftp.esat.kuleuven.be/pub/SISTA/ida/reports/01-05.pdf>
- [11] GLOROT, X.; BENGIO, Y.: Understanding the difficulty of training deep feedforward neural networks. 2010, [Online], [cit. 2015-12-17].
URL http://machinelearning.wustl.edu/mlpapers/paper_files/AISTATS2010_GlorotB10.pdf

- [12] GOMBER, P.; ARNDT, B.; LUTAT, M.; aj.: High-Frequency Trading. 2011, [Online], [cit. 2015-12-17].
 URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1858626
- [13] HARDT, J.; HERKE, M.; BRIAN, T.; aj.: Multiple Imputation of Missing Data: A Simulation Study on a Binary Response. [Online], [cit. 2016-03-14].
 URL <http://www.scirp.org/journal/PaperInformation.aspx?PaperID=37663>
- [14] HINTON, G.; Deng, L.; Yu, D.; aj.: Deep Neural Networks for Acoustic Modeling in Speech Recognition. 2012, [Online], [cit. 2015-12-17].
 URL <http://static.googleusercontent.com/media/research.google.com/cs/pubs/archive/38131.pdf>
- [15] HIRSHLEIFER, D.; SHUMWAY, T.: Good Day Sunshine: Stock Returns and the Weather. [Online], [cit. 2016-03-04].
 URL <http://www-personal.umich.edu/~shumway/papers.dir/weather.pdf>
- [16] HUF, P.: *Machine Learning Strategies in Electronic Trading*. Diplomová práce, FIT VUT v Brně, 2014, [Online], [cit. 2015-12-17].
 URL <http://www.fit.vutbr.cz/study/DP/BP.php?id=16764&y=2013>
- [17] ICE: White Sugar Futures. [Online], [cit. 2015-12-17].
 URL <https://www.theice.com/products/37089080/White-Sugar-Futures>
- [18] KŘÍŽ, J.: *Algoritmické obchodování na burze s využitím dat z Twitteru*. Diplomová práce, FIT VUT v Brně, 2015, [Online], [cit. 2015-12-17].
 URL <http://www.fit.vutbr.cz/study/DP/DP.php?id=17799&y=2014>
- [19] Learning, U. F.; Learning, D.: Neural Networks. [Online], [cit. 2016-05-15].
 URL http://ufldl.stanford.edu/wiki/index.php/Neural_Networks
- [20] LECUN, Y.; BOTTOU, L.; ORR, G.; aj.: Efficient BackProp. 1998, [Online], [cit. 2015-12-17].
 URL <http://yann.lecun.com/exdb/publis/pdf/lecun-98b.pdf>
- [21] MACEY, J.; O'HARA, M.; POMPILIO, D.: Down and Out in the Stock Market: The Law and Finance of the Delisting Process. 2004, [Online], [cit. 2015-12-17].
 URL <http://www.haas.berkeley.edu/groups/finance/delistings%20-%20Mar04%20draft.pdf>
- [22] MENNE, M.; DURRE, I.; VOSSE, R.; aj.: An Overview of the Global Historical Climatology Network-Daily Database. 2012, [Online], [cit. 2015-12-17].
 URL <http://journals.ametsoc.org/doi/pdf/10.1175/JTECH-D-11-00103.1>
- [23] MIKOLOV, T.: *Statistical language models based on neural networks*. Dizertační práce, FIT VUT v Brně, 2012.
- [24] MIKULENČÁK, R.: *Predikce kursů pro obchodování na akciových trzích*. Diplomová práce, FIT VUT v Brně, 2015, [Online], [cit. 2015-12-17].
 URL <http://www.fit.vutbr.cz/study/DP/DP.php?id=17801&y=2014>

- [25] MOAT, H. S.; CURME, C.; AVAKIAN, A.; aj.: Quantifying Wikipedia Usage Patterns Before Stock Market Moves. [Online], [cit. 2016-03-04].
 URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2263897
- [26] MONNER, D.; REGGIA, J.: A generalized LSTM-like training algorithm for second-order recurrent neural networks. [Online], [cit. 2016-05-15].
 URL <http://www.overcomplete.net/papers/nm2012.pdf>
- [27] NASSIRTOUSSI, A. K.; AGHABOZORGI, S.; WAH, T. Y.; aj.: Text mining for market prediction: A systematic review. *Expert Systems with Applications*, ročník 41, č. 16, 2014.
- [28] National Centers for Environmental Information: About Us. [Online], [cit. 2015-12-17].
 URL <https://www.ncdc.noaa.gov/about>
- [29] NYSE: NYSE overview statistics. [Online], [cit. 2015-12-17].
 URL http://www.nysedata.com/nysedata/asp/factbook/viewer_edition.asp?mode=table&key=268&category=14
- [30] NYSE: Suspension and Delisting. [Online], [cit. 2015-12-17].
 URL http://nysemanual.nyse.com/LCMTTools/PlatformViewer.asp?selectednode=chp_1_9&manual=%2Flcm%2Fsections%2Flcm-sections%2F
- [31] OGASAWARA, E.; MARTINEZ, L. C.; de OLIVIERA, D.; aj.: Adaptive Normalization: A Novel Data Normalization Approach for Non-Stationary Time Series. 2010, [Online], [cit. 2016-03-14].
 URL <http://homepages.dcc.ufmg.br/~glpappa/papers/Ogasawaraetal-2010-IJCNN.pdf>
- [32] Olah, C.: Understanding LSTM Networks. [Online], [cit. 2016-03-04].
 URL <http://colah.github.io/posts/2015-08-Understanding-LSTMs>
- [33] Penize.cz: Akcie. [Online], [cit. 2015-12-17].
 URL <http://www.penize.cz/akcie>
- [34] PHAM, V.; BLUCHE, T.; KERMORVANT, C.: Dropout improves Recurrent Neural Networks for Handwriting Recognition. 2013, [Online], [cit. 2016-03-25].
 URL <http://arxiv.org/pdf/1312.4569.pdf>
- [35] Portál veřejné správy: Zákon České národní rady ze dne 20. listopadu 1992 o daních z příjmů. [Online], [cit. 2015-12-17].
 URL <https://portal.gov.cz/app/zakony/download?idBiblio=40374&nre=586~2F1992~20Sb.&ft=pdf>
- [36] PREIS, T.; MOAT, H. S.; STANLEY, H. E.: Quantifying Trading Behavior in Financial Markets Using Google Trends. [Online], [cit. 2016-03-04].
 URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2260189
- [37] REESE, A.: Geography Review. [Online], [cit. 2016-03-04].
 URL http://diggingintohistory.blogspot.cz/2015_08_01_archive.html

- [38] ROSS, S.: Survivorship Bias in Performance Studies. [Online], [cit. 2015-12-17].
 URL <http://www.cfapubs.org/doi/pdf/10.2469/cp.v1994.n9.9>
- [39] SCHMIDHUBER, J.: Tutorial on LSTM Recurrent Nets. [Online], [cit. 2016-05-15].
 URL <http://people.idsia.ch/~juergen/lstm2003tutorial.pdf>
- [40] SMITH, L. I.: A tutorial on Principal Components Analysis. 2002, [Online], [cit. 2016-03-14].
 URL http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
- [41] SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; aj.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. [Online], [cit. 2016-03-14].
 URL <https://www.cs.toronto.edu/~hinton/absps/JMLRdropout.pdf>
- [42] StockCharts: Bollinger Bands. [Online], [cit. 2016-05-15].
 URL http://stockcharts.com/school/doku.php?id=chart_school:technical_indicators:bollinger_bands
- [43] StockCharts: Historical Price Data is Adjusted for Splits, Dividends and Distributions. [Online], [cit. 2015-12-17].
 URL http://stockcharts.com/docs/doku.php?id=policies:adjusted_data
- [44] StockCharts: Introduction to Technical Indicators and Oscillators. [Online], [cit. 2016-05-15].
 URL http://stockcharts.com/school/doku.php?id=chart_school:technical_indicators:introduction_to_technical_indicators_and_oscillators
- [45] StockCharts: William %R. [Online], [cit. 2016-05-15].
 URL http://stockcharts.com/school/doku.php?id=chart_school:technical_indicators:williams_r
- [46] WALCZAK, S.: An Empirical Analysis of Data Requirements for Financial Forecasting with Neural Networks. 2001, [Online], [cit. 2016-03-21].
 URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.6904&rep=rep1&type=pdf>
- [47] WAYMAN, J. C.: Multiple Imputation For Missing Data: What Is It And How Can I Use It? 2003, [Online], [cit. 2016-03-21].
 URL http://www.csos.jhu.edu/contact/staff/jwayman_pub/wayman_multimp_aera2003.pdf
- [48] YUAN, Y. C.: Multiple Imputation for Missing Data: Concepts and New Development. [Online], [cit. 2016-03-21].
 URL <http://www.ats.ucla.edu/stat/sas/library/multipleimputation.pdf>
- [49] ZAREMBA, W.; SUTSKEVER, I.; VINYALS, O.: Recurrent Neural Network Regularization. 2015, [Online], [cit. 2016-03-25].
 URL <http://arxiv.org/pdf/1409.2329.pdf>
- [50] ÖZTÜRK, F.; ÖZEN, F.: A New License Plate Recognition System Based on Probabilistic Neural Networks. *Procedia technology*, 2012.

- [51] ČERNOCKÝ, J.: Temporal processing for feature extraction in speech recognition. 2013, [Online], [cit. 2016-03-14].
URL <http://www.fit.vutbr.cz/~cernocky/publi/2003/vutium.pdf>

Přílohy

Seznam příloh

A Obsah CD	98
B Manuál	99
B.1 Prekvizity	99
B.2 Konfigurace	99
B.3 Výpočet systému	99
C Plakát	101

Příloha A

Obsah CD

- /Code – obsahuje veškeré použité zdrojové kódy.
- /Data – obsahuje pouze soubor *FullData* s daty dle výsledné konfigurace.
- /MI – zdrojový kód použitý pro doplnění chybějících dat metodou multiple imputation.

Příloha B

Manuál

Zde se nachází popis zprovoznění vyvíjeného systému a jeho používání.

B.1 Prekvizity

Veškerý kód je napsán v jazyce C# jako konzolová aplikace. Pro kompliaci je potřeba mít nainstalované Visual Studio. Některé knihovny jsou dodány přímo a jsou součástí projektu. Další knihovny se automaticky stáhnou při kompliaci pomocí balíčkovacího systému NuGet. Dále je vyžadován .NET Framework 4.5. Pro doplňování metodou multiple imputation je potřeba nainstalovat software R¹.

B.2 Konfigurace

Základní konfigurace je prováděna pomocí třídy *Config*. Tato třída je nastavena při spuštění programu přímo v komponentě *Manager*. Zásadní je nastavit cestu k datům. Tato cesta se nastavuje v proměnné *DataDirectory*. Pro stahování dat je třeba zadat API klíč od Quandl do proměnné *QuandlApiKey*. Ten lze získat bezplatnou registrací na <https://www.quandl.com/>. Dále je dobré zadat do proměnné *GoogleCookies* cookies uložené v prohlížeči pro google.com. Tyto cookies se dají získat přihlášením do Googlu a načtením stránky <https://www.google.com/trends/>. Cookies se do proměnné uloží ve formátu *Key=Value;Key2=Value2;Key3=Value3;...*. Zadání cookies umožní zvětšit limit pro stažení dat z Google Trends. Další položky je vhodné ponechat v původním nastavení.

B.3 Výpočet systému

Systém si automaticky stáhne veškerá potřebná data do specifikované složky. Po stažení dat následuje jejich zpracování. Každý zdroj dat má svoji vlastní složku. V této složce je kromě těchto dat uložen soubor *_Serialized.bin*. Tento soubor obsahuje už zpracovaná data v binárním formátu a slouží jako první stupeň mezipaměti. Pokud dojde ke změně například čistícího procesu pro určitý zdroj dat, je potřeba tento soubor smazat aby se data zpracovala znova. V prvním kroku tedy systém prochází veškeré zdroje dat a data budou zpracovává, nebo je načítá již zpracovaná v binárním formátu ze souboru *_Serialized.bin*.

¹<https://www.r-project.org/>

Po nasbírání všech potřebných data tato data spojí do jedné velké matice, vyčistí je a vytvoří soubor *FullData* ve složce *TradingData*. Tento soubor slouží jako další úroveň mezipaměti. Pokud je tento soubor přítomný, nedochází ke sběru dat, ale systém automaticky začne načtením souboru *FullData*. Proto je na CD tento soubor, aby uživatel nemusel stahovat veškerá data, pokud nechce.

V dalším kroku tedy systém načte soubor *FullData* a vytvoří soubory pro neuronovou síť. Tento krok provádí třída *RnnTable*. Zde také dochází k normalizaci, tvorbě vstupu a výstupu, výběru datových řad apod. Soubory vytvořené v tomto kroku slouží opět jako další stupeň mezipaměti.

Následně je řízení předáno do třídy *RnnManager*, kde je prováděno trénování neuronové sítě a její predikce na testovaných datech. Výstupní soubory obsahující predikce jsou dalším stupněm mezipaměti.

V dalším kroku je provedena denormalizace predikovaných dat. Výstupem jsou soubory s denormalizovanými daty a slouží opět jako další stupeň mezipaměti.

Posledním krokem je simulace obchodního systému. Tuto simulaci provádí třída *TradingSystem*. Výstupem je soubor *Trades*, který obsahuje vývoj účtu na trénovacích a validačních datech (případně i testovacích).

Při testování je tedy potřeba vždy smazat vytvořené soubory až do úrovně, od které má být výpočet prováděn.

Příloha C

Plakát

Fundamentální analýza
v elektronickém obchodování

DEMOGRAPHY

Unemployment

- ➡ Velké množství dat
- ➡ LSTM neuronová síť
- ➡ Dlouhodobá predikce
- ➡ Kvalitní obchodní strategie

Začněte obchodovat akcie ještě dnes!

autor
vedoucí

Petr Huf
Jan Černocký