FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Zadání bakalářské práce

Název:	Predikce signálu na finančních trzích pomocí analýzy časových řad
Student:	Bohumil Miláček
Vedoucí:	Ing. Stanislav Kuznetsov
Studijní program:	Informatika
Obor / specializace:	Znalostní inženýrství
Katedra:	Katedra aplikované matematiky
Platnost zadání:	do konce letního semestru 2022/2023

Pokyny pro vypracování

Cílem práce je predikce signálu na finančních trzích na základě analýzy časových řad vývoje určitých akcií. Abychom mohli signály generovat, musíme provést transformaci tzv. tickových dat na časovou řadu. Na základě správně zvolené interpretace časové řady a algoritmu data miningu bychom chtěli generovat signály pro nákup/prodej akcií.

1. Proveďte rešerši a popište nejpoužívanější způsoby převodu tickových dat na časovou řadu.
2. Vyberte jeden způsob převodu a popište důvod jeho výběru.
3. Proveďte preprocessing burzovních dat a vyberte správné metriky.
4. Vytvořte modely a proveďte experimenty s ohledem na vybrané metriky.
5. Ukažte na příkladech práci hotových modelů.

Elektronicky schválil/a Ing. Karel Klouda, Ph.D. dne 1. března 2021 v Praze.

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Bakalářská práce

Predikce signálu na finančních trzích pomocí analýzy časových řad

Bohumil Miláček

Katedra aplikované matematiky
Vedoucí práce: Ing. Stanislav Kuznetsov

13. května 2021

Poděkování

Rád bych poděkoval svému vedoucímu práce Ing. Stanislavu Kuznecovovi za vedení a konzultace. Dále Milanu Hemzalovi za připomínky a rady k dokončení této práce. V poslední řadě velké díky patří rodině a všem kteří mě podporovali.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 2373 odst. 2 zákona č. 89/2012 Sb., občanský zákoník, ve znění pozdějších předpisů, tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou a veškeré jejich dokumentace (dále souhrnně jen „Dílo"), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu) licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 13. května 2021 . . . . . .

Odkaz na tuto práci

Miláček, Bohumil. *Predikce signálu na finančních trzích pomocí analýzy časových řad*. Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2021.

Abstrakt

Práce se zaměřuje na analýzu nejvhodnějších typů finančních časových řad pro Machine learning na predikci buy/sell signálů. Data vybraná pro tuto práci byly akcie, konkrétně TSLA, AAPL, MSFT. Byla vzata jejich Tick data, agregována na time, tick, volume, dollar a Renko svíčky. Na jednotlivých svíčkách byl proveden labeling metodou CTL, která značí trend. Pomocí triviální strategie byl nasimulován trading na jednotlivých typech svíček. Pro vyhodnocení byly zvoleny klasifikátory Random forest a Neuronová síť, na kterých byla data natrénována a vyhodnocena vhodnost pro modely Machine learningu na základě výsledků předpovídání trendu na svíčkách a výsledků simulování tradingu na predikovaném buy/sell signálu. Výsledky ukázaly, že nejvhodnějším typem svíček jsou Renko.

Klíčová slova finanční trh, analýza transformací tick dat, tick data, časová řada, data labeling, predikce signálu, machine learning, Python

Abstract

Work focuses on analysis of most suitable financial time series for Machine learning to predict buy/sell signals. Data selected for this thesis were stocks, namely TSLA, AAPL, MSFT. Their Tick data were taken and aggregated to time, tick, volume, dollar and Renko bars. On all bars a labeling was done using CTL method, which marks trend. A trivial strategy was used to simulate trading on all of the bars. As classifiers for evaluation a Random forest and Neural network were used, on which were the data trained and evaluated suitability for Machine learning models based on results of trend prediction on bars and results of simulated trading on predicted buy/sell signal. Results showed the most suitable bars are Renko.

Keywords financial market, tick data transformation analysis, tick data, time series, data labeling, signal prediction, machine learning, Python

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Obchodování na finančních trzích za účelem zisku, neboli nakoupit levně a prodat draze, je každodenní záležitostí mnoha lidí i firem a v poslední době i strojů. Tato činnost se stává stále rozšířenější a populárnější, stále více jedinců ji začíná provádět, vznikají firmy které se zabývají pouze touto problematikou.

Obchoduje se s různými aktivy jako jsou akcie, komodity, opce, kryptoměny, měny. Všechny tyto věci mají jedno společné, během času mění svou hodnotu. Tuto vlastnost můžeme využít právě k zmiňovanému obchodování. Otázkou je, kdy nakoupit a kdy prodat. Za tímto účelem probíhají nejrůznější analýzy. Matematici dělají statistické a pravděpodobnostní odhady aby zjistili jak se bude hodnota pohybovat.

Myšlenka počítačového programu, který sám obchoduje, je nejen fascinující, ale v dnešní době se stala už ideálním přístupem k obchodování. Od přelomu století se Algoritmický trading stal na finančních trzích dominantní. V rozmezí 5–10 let nazpět 70–80 % obchodů na americké burze, oproti roku 2003 kdy se jednalo o pouze okolo 15 %. Podobně je tomu i na Forexu [1].

V dnešní době má velký rozmach umělá inteligence, konkrétně její podčást – Machine Learning (ML), ve kterém jde o extrahování znalostí z dat a učení se na datech z minulosti. ML se datoví inženýři a analytici snaží aplikovat na obchodování. Metody jako deep learning, rozhodovací stromy, rekurentní neuronové sítě, LSTM, regrese, Fourierovy transformace a kombinace těchto metod. Aktuálně jsou stále více aplikovány na analýzy časových řad.

Problém finančních časových řad je, že jejich hodnoty je velmi obtížné odhadnout. Nejsou zpravidla periodické, nemají lehce odhadnutelnou korelaci s faktory na základě kterých se jejich hodnota pohybuje, jichž je nepřeberné množství. Ve své podstatě jde hlavně o dva faktory a sice poptávka a nabídka. S velkou poptávkou roste cena, protože lidé jsou ochotni zaplatit více za produkt. Naopak s velkou nabídkou cena klesá. Pak tady ale máme další vlivy, jež ovlivňují i samotnou poptávku a nabídku. Například média – podle toho co o produktu napíší, je více či méně žádán, ekonomický stav – růst nebo krize,

politické situace, dokonce i samotná komunita lidí.

Je veliká snaha o to vytvořit co nejlepší model takový, že bude schopen předpovídat trendy a pohyby finančních časových řad. Kdo takový model vytvoří získá na trhu obrovskou výhodu. Před sestrojením takového modelu je potřeba udělat analýzu trhu a datech na něm, aby byla data pro model co nejvhodnější, neboli aby obsahovala co nejvíce informací, ze kterých se může učit. Tomu se budu věnovat v první části a to je hlavním tématem a výstupem bakalářské práce. V druhé části naučím model na datech, zjistím, vhodnost různých typů dat pro model na základě několika metrik a provedu několik experimentů. Jedná se o velmi aktuální téma, právě proto se jím budu v práci zabývat.

Kapitola 1

Cíl práce

V rešerši se budu zabývat metodami transformací burzovních Tick-by-Tick, která obsahují informace o uskutečněných obchodech na burze, na reprezentaci OHLC. Konkrétně na časové, tickové, volume, dolarové svíčky a renko. U každých svíček rozepíši jejich vlastnosti, výhody či nevýhody.

Dále popíši metody strojového labelingu finančních časových řad, jednu z nich si vyberu a použiji na labeling buy/sell signálů všech typů svíček.

Na svíčkách s labelama natrénuji modely ML, vyhodnotím na základě několika metrik a simulování tradingu vhodnost či nevhodnost různých svíček pro strojové učení.

V závěru shrnu veškeré dosažené poznatky ohledně vhodnosti určitých svíček pro Machine learning oproti jiným pro generování signálů na burze.

Část I

Teoretická část

Kapitola 2

Stroje a finanční trhy

V počátcích obchodování na burze bez počítačů bylo nákupním a prodejním signálem zvednutí ruky a výkřiky na pevně daném místě burzy. Se vzestupem a rychlým vývojem počítačů se převedlo obchodování do virtuálního světa počítačů [2].

Jako další přišla éra automatického a algoritmického obchodování, kdy počítače obchodují samy, když dojde ke splnění určitých předpokladů, analýzou trhu v reálném čase. Ale algoritmické obchodování není perfektní. Tímto způsobem říkáme, že se má provést akce A, za předpokladu že nastane B. Například A - nakoupit, když nějaká akce B - klesne. To je automatické obchodování na základě naprogramovaných podmínek člověkem, ale to může selhat jakmile nastanou neočekávané události [2].

V tuto chvíli přichází na scénu strojové učení, které jde o krok dál. Můžeme jej natrénovat na tisícovkách různých typů dat, aby rozumělo burze, proč a jak se pohybuje [2].

2.1 Výhody a nevýhody

Přestože se jedná o velice složitý problém, mnoho lidí a firem se snaží o implementování botů na finanční trhy. To přináší hned několik výhod, ty hlavní například následující:

- Ušetření času a peněz tradera, za kterého obchoduje bot
- Zvýšená produktivita oproti traderovi
- Analýza obrovského množství dat, které ovlivňují aktivum, díky velkému výpočetnímu výkonu počítačů oproti lidem, kteří by toto nezvládli
- Obchodování bez emocí

Oproti tomu velkou nevýhodou jsou obrovské náklady na vývoj.

2.2 Historie

Mnoho problémů spojených s oblastí umělé inteligence je probíraných již desítky let dozadu. Burza má již nějakou historii s AI. V 60 letech 19. století se mnoho výzkumů soustředilo na Bayesovskou statistiku, metody hojně používané ve strojovém učení, aplikované též na předpovídání burzy. 80 léta zaznamenala rozmach expertních systémů. V této době již 2/3 Fortune 1000 (1000 největších US firem podle jejich výnosu) mělo ve vývoji alespoň jeden AI projekt [3].

Mnoho výzkumu AI provedeného v 50 a 60 letech 19. století, se nezaměřovalo na oblast financí, avšak hodně matematiky z výzkumů se používá dodnes. Užívání pokročilé matematiky ve financích začalo společně s vydáním práce *Louise Bacheliera Théorie de la Spéculation (Theory of Speculation)*. Jeho práce je považována jako jedna z prvních, která se zabývá výzkumem použití matematiky na vyhodnocování akcií. Toto byl začátek pro používání statistických modelů k primitivnímu AI ve finančním světě [3].

Dalším mezníkem byla práce Roberta Schlaifera, který zpopularizoval Bayesovskou statistiku svou prací *Bayesian Decision Theory*. Ta zakládala na používání Bayesovské statistiky k provádění informovaných rozhodnutí na základě pravděpodobností a práce *Probability and Statistics for Business Decision* ještě více téma „Matematika v byznysu“ zpopularizovala [3].

Během 80 let se objevilo mnoho praktik používaných k tvorbě AI pro oblast financí, jako například umělé neuronové sítě nebo fuzzy systémy. Ale mnoho zájmu bylo věnováno systémům založených na znalostech tzv. znalostním systémům. Například americká společnost DuPont si nechala vyvinout okolo 100 takových systémů, které jim pomohly ušetřit miliony dolarů ročně. Jeden z prvních takových programů, který hypoteticky předpovídal burzu, byl Protrader Expert system vytvořen K. C. Chenem, jehož hlavní funkce byly monitorování trhu, rozhodnout optimální investiční strategii, provést transakce a adekvátně upravit bázi znalostí za pomocí machine learning mechanismu. Expertní systémy ale neuspěly v oboru financí a to z několika důvodů. Byly až moc komplikované anebo se nesetkaly s požadavky klientů [3].

2.3 Stroje na burze v současnosti

Problém předpovídání pohybu finanční časové řady je velmi složitý a to z důvodu obrovského množství faktorů na kterém hodnota řady závisí. Například co napíší média. Od toho se hodnota hodně odvíjí. Mohli bychom tedy napsat bota, který si přečte články o dané věci, která je předmětem časové řady a vyhodnotí podle kontextu jestli se jedná o pozitivní, neutrální nebo negativní zprávu a v jak moc velkém měřítku. Sestrojit takový model je ale nesmírně složitá věc, protože stroj může vyhodnotit zprávu negativně i když je myšlena pozitivně a naopak. To vše kvůli složitosti jazyka. Například „Není to vůbec špatné“ by mohl vyhodnotit negativně, přestože se jedná o celkem kladnou zprávu.

AI je již v dnešní době nasazena na trh ve velkém měřítku. Malý trader soutěží na trhu s AI, HFT, AMM a super-počítačemi v hodnotách stovek milionů dolarů, které zpravidla používají zmíněné metody [4]. Toto značně trading znemožňuje a spíše navádí na dlouhodobý trading a investování.

AI systémy obchodující na burze se rychle zlepšují. V Červnu roku 2020 výzkumníci z Univerzity v Cagliari oznámili výsledky projektu, který použil konvoluční neuronovou síť (CNN), jež řídila buy and hold strategii. Byla natrénovaná na historických datech indexu S&P 500 a dokázala předpovídat trh v reálném čase s 50% přesností [2].

I když je pořád brzo na to říci, zda ML je jasná budoucnost pro obchodování na burze, mnoho ukazatelů tomu napovídá. Už existují systémy, které dokáží předvést lepší než průměrné výsledky. Na druhou stranu pořád nedokáží plně zreprodukovat intuici člověka. Co je jasné je, že ML obchodování má navrch oproti obchodování algoritmy, který je limitovaný dovednostmi svého uživatele [2].

Kapitola 3

Porozumění problematice a datům

V této sekci se zabývám s jakými daty budu pracovat a jejich popis. Dále se budu věnovat metodám převodu neagregovaných burzovních dat na strukturovanější data, které detailněji popíšu. Výzkumníci na finančních trzích věří, že například cena či objem obchodovaného aktiva může být použit jako indikátor budoucích změn v ceně aktiva a proto může poskytnout informaci o výnosnosti [5]. Z toho důvodu se na toto v následující sekci zaměřím. V poslední části se zaměřím na metody labelingu časových řad a metriky pro vyhodnocení klasifikátoru.

3.1 Data

Data kterými se v práci zabývám se jmenují high-frequency data [6, 7]. Jedná se o časové řady, které mají velmi mnoho záznamů a dochází k jejich tvoření ve vysoké frekvenci a to z důvodu, že dochází k obchodování obrovského množství cenných papíru v rámci milionů až miliard za den. Jsou primárně používány na finančních trzích a především k analýze akciového trhu. V tradingu je velmi důležité jakou reprezentaci dat použijeme, jelikož nám může odhalit různé vzory chování časové řady a trendy [8].

3.1.1 Tick data

Jako tick data, někdy také tick-by-tick data, označujeme neagregované záznamy všech hodnot cenného papíru, které vzniknou v rámci jednoho dne. Mají nejmenší granularitu ze všech high frequency dat [6]. Jeden záznam označujeme jako Tick. Ten se vytvoří vždy při každém obchodu akcie spolu s informacemi čas, datum, za kolik se obchodovala a množství [9].

3.1.2 OHLC

Jedná se o časovou řadu poskládanou z takzvaných svíček, kde každá svíčka obsahuje hodnotu:

- Open – hodnota, při které byla svíčka otevřena
- High – hodnota, která byla ve svíčce největší
- Low – hodnota, která byla ve svíčce nejnižší
- Close – hodnota, při které byla svíčka zavřena

Prvním krokem k vytvoření skvělého modelu je agregovat data do vhodného formátu pro další analýzu. Svíčky obsahují nejzákladnější informace o pohybech ceny nějakého finančního aktiva. Svíčky jsou typický formát vstupních dat for trénování a testování ML modelu. Lze si snadno představit, že způsob předzpracování a agregace dat může mít velký vliv na výsledný model [10].

3.2 Převod tick dat na OHLC

Přestože se zdá intuitivní tick data převést na časovou řadu pomocí fixních časových intervalů minuty/hodiny/dny... Nemusí to být úplně tak pravda. Informace v tick datech nemusí být distribuované rovnoměrným rozdělením, ale může být například v různých částech dne různá aktivita a tudíž tento fakt například podchytit, což přesně stanovené časové intervaly nemusí. Budeme se snažit, aby ideálně každá svíčka obsahovala stejně informací jako ostatní svíčky [10]. Podíváme se tedy na pár metod převodu tickových dat na OHLC. V následujících podsekcích si popíšeme tyto metody, analyzujeme si jejich vlastnosti a podíváme se na jejich vhodnost resp. nevhodnost v určitých situacích.

3.2.1 Time-based bars

V tomto případě se převádí tick data pomocí pevně stanoveného časového intervalu. Jedná se o základní a nejčastější metodu převodu tick dat na časové řady. Vezme se interval například 5 minut, ticky se rozdělí na množiny ticků po 5ti minutách a z každé množiny sevytvoří svíčka. Každý den bude počet svíček stejný. Zpravidla se používají menší intervaly v rámci minut. Když dochází k pár transakcím, zdá se že časové svíčky ukazují více informací [11].

3.2.2 Tick-based bars

Tento typ svíček je vytvářen na základě pevně stanoveného počtu N ticků. Data se rozdělí na množiny, kde každá bude obsahovat N ticků, z každé množiny pak vznikne jedna svíčka. Toto způsobí lepší zachycení informací,

když dochází k více obchodům. Tickové svíčky tedy zachytí více informací než časové svíčky, když dochází k velké aktivitě. Tyto informace zahrnují například cenové změny, konsolidaci, menší změny v ceně aktiva. Pokud v první minutě dochází k velkým změnám a obchodům na aktivu, tickové svíčky tuto událost odhalí mnohem dříve než časové a může to být využito ku prospěchu obchodování [11].

3.2.3 Volume-based bars

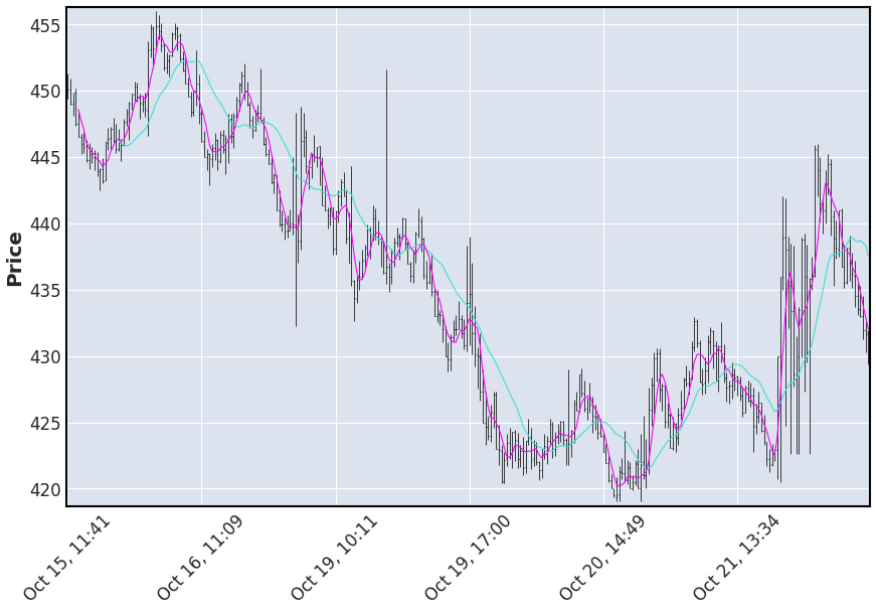
Jednou z nevýhod tickových svíček je, že ne každý obchod je rovnocenný. Ticky totiž obsahují informaci „Volume“, která nám říká kolik se aktiva obchodovalo. Z hlediska tickových svíček tedy 10 obchodů s množstvím 100 bude stejné jako 10 obchodů s množstvím 10 000 aktiva, z toho důvodu by mohly volume svíčky poskytnout ještě větší náhled, jelikož přímo značí jak moc se aktivum obchoduje. Další motivací je tedy rozdělit data podle pevně stanoveného množství N objemu obchodovaného aktiva [10]. Podobně jako u tickových svíček můžeme pozorovat jak rychle se trh pohybuje [12].

3.2.4 Dollar-based bars

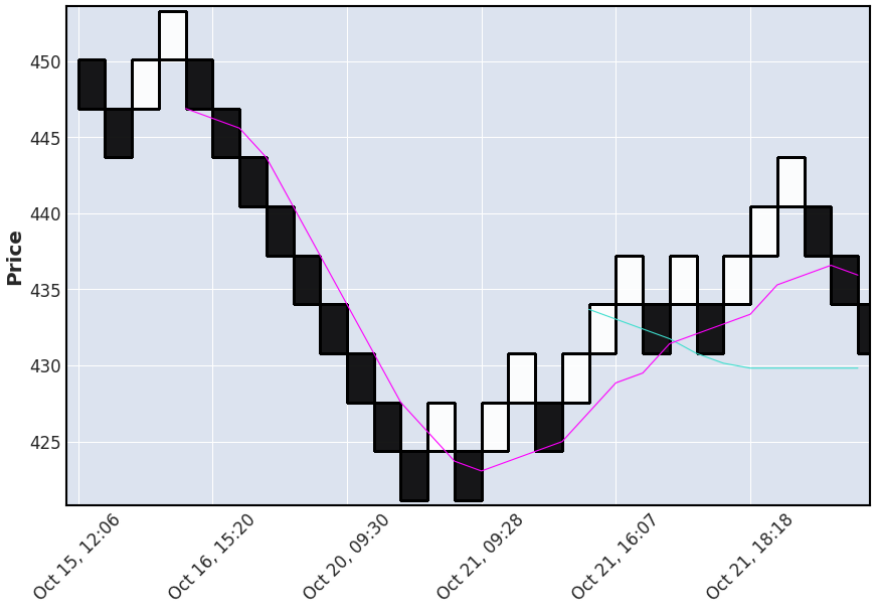
Pokud jako fixní hodnotu, podle které vytváříme z dat svíčky, vezmeme cenu za kterou se aktivum obchodovalo nazýváme je Dolarovými svíčkami. Ve své podstatě jsou nejlepší na statistickou analýzu, protože jejich návratnost je nejblíže normálnímu rozdělení na rozdíl od ostatních svíček [13].

3.2.5 Renko

Renko je typ reprezentace, který se tvoří se na základě pohybu ceny. Nová Renko svíčka je vytvořena když se cena pohne o předem pevně stanovenou hodnotu – „velikost boxu“. Z toho plyne, že hodnoty High a Low se překryjí s hodnotami Open a Close, zůstanou nám tedy pouze hodnoty otevření a uzavření. Některé svíčky trvá vytvořit déle než jiné, z důvodu, že se cena může delší dobu pohybovat nahoru a dolu a nepřekročí velikost boxu, aby se vytvořila svíčka. Renko tedy vyfiltruje pohyby, které jsou menší než velikost boxu, čímž odstraní šum, což poté dává traderům lepší pohled na trendy akcie [14].



(a) OHLC



(b) Renko

Obrázek 3.1: Porovnání klasických OHLC (a) a Renko (b) na stejném časovém intervalu akcie TSLA. Zdroj autor.

3.3 Labeling dat

Ve strojovém učení data labeling je proces, při kterém analyzujeme data a přidáváme jeden či více informačních příznaků (label), abychom ML modelu poskytly informace, na základě kterých se učí [15]. Když chceme používat supervizované strojové učení na předpovídání, hned po předzpracování je klíčové a velmi důležité provést správný labeling dat pro co nejlepší výsledek. Jak provést správný labeling, je v poslední době značně populárnějším tématem. Aktuálně používané metody labelingu finančních časových řad převážně pracují na principu porovnávání aktuálních dat s daty, které se nacházejí v blízké budoucnosti. Avšak časové řady na finančních trzích, mívají nelineární chování s náhodností v krátkodobějším horizontu. Proto tyto metody nezachytí trend řady [16]. V následujících podsekcích si představíme některé tyto metody a dostaneme se od těch naivních až po sofistikovanější.

3.3.1 Labeling fixním časovým intervalem

Labeling pomocí fixního časového intervalu je velmi triviální metoda. Data labeling je prováděn na základě návratnosti v budoucnu pomocí fixního časového horizontu. Labely jsou přiřazeny na základě Stop-Loss a Take-Profit prahu [17]. Toto je zachyceno v tabulce (3.1), kde X_t je hodnota aktiva v čase t a parametr n je zvolený fixní časový horizont do budoucnosti.

Tabulka 3.1: Data labeling metodou s fixním časovým intervalem

Label	1	0	-1
Podmínka	$X_{t+n} > X_t$	$X_{t+n} = X_t$	$X_{t+n} < X_t$
Návratnost	Kladná	Nulová	Záporná

Prahy jsou fixní, ale volatilita není, což znamená, že prahy mohou být moc daleko od sebe, nebo naopak moc blízko [17].

3.3.2 Triple barrier

Pro lepší simulaci trading strategie je vhodné nastavovat práh na základě volatility aktiva. Na základě zpětného pozorování n-periodického okna, jsou zadefinovány 2 prahy. Horní práh B_u a dolní práh B_d. Jejich výpočet je v následující rovnici (3.1), kde p značí aktuální cenu, v volatilitu. Pro výpočet volatility v existuje několik způsobů. Například směrodatnou odchylkou σ (3.2), Average True Range (3.4), nebo Beta β (3.5).

$$B_u = p + v \qquad B_d = p - v \tag{3.1}$$

$$\sigma = \sqrt{var(X)} \tag{3.2}$$

$$TR = \max(h - l, |h - c_p|, |l - c_p|) \tag{3.3}$$

$$ATR = \frac{1}{n}\sum_{i=1}^{n} TR_i \tag{3.4}$$

$$\beta = \frac{cov(R_a)(R_m)}{var(R_m)} \tag{3.5}$$

V rovnici (3.3) h značí nejvyšší hodnotu akcie během dne, l nejnižší hodnotu akcie během dne, c_p je uzavírací hodnota akcie předchozího dne(vzorec se mění podle typu OHLC). V rovnici (3.5) $cov(R_a, R_m)$ značí kovarianci návratnosti akcie R_a a celkovou návratnost trhu R_m, neboli jak návratnosti akcie souvisí s návratností trhu. Variance $var(R_m)$ jak moc se trh pohybuje od svého průměru [18].

Toto je mnohem lepší metoda, ale nereálná pro ty, kteří chtějí držet pozici dokud cena nepřekročí práh. Řešení je vytvoření 3. prahu, který je vertikální, který omezí pohled metody dopředu do budoucnosti. Tuto metodu se 3 prahy nazýváme Triple barrier, její princip spočívá v tom, kterého prahu se cena dotkne jako první. Data labeling je zachycen v tabulce (3.2), kde X_t značí hodnotu akcie v čase t, n je časový úsek jak daleko do budoucnosti hodnoty akcie koukáme, B_u a B_d jsou horní a dolní práh.

Tabulka 3.2: Data labeling metodou s Triple barier

Label	1	0	-1
Podmínka	$X_{t+n} > B_u$	$B_d < X_{t+n} < B_u$	$X_{t+n} < B_d$
Návratnost	Kladná	Kladná/Nulová/Záporná	Záporná

Obrázek 3.2: Metoda Triple barier na akcii AAPL (NASDAQ). a – počáteční datum, b – Stop-Loss práh, c – Take-Profit práh, d – počáteční datum + počet dní jak dlouho plánujeme akcii držet. Zdroj [19].

Parametr n, v n-periodickém okně metody, je pouhý hyperparamter, který můžeme ladit podle našeho úsudku, jak moc konzervativní chceme v tradingu být [20].

3.3.3 Continuous trend labeling

Když trh následuje plynulé trendy, je rozdělen na stoupající a klesající trh. Investoři by měli nakoupit a držet aktivum ve stoupajícím trhu, ale držet short pozici s short mechanismem. Pokud žádný short mechanismus není, investoři by měli prodat aktivum v klesajícím trhu. Jejich pozice by se neměla změnit, dokud se předpověď trendu aktiva nemá změnit [21]. V práci zabývající se finančními trhy a data labeling [16] je uveden způsob a algoritmus, jak od sebe odlišit jednotlivé plynule klesající a plynule stoupající trendy trhu následující definicí.

Zprvu se vezmou peak body a body trough z historických dat a položí se do vektroů h a l, kde t značí počet peak bodů a m značí počet trough bodů v rovnici (3.6). Byl použit index TD, který reprezentuje výpočet stupeň trendu nějaké časové řady. Výsledek TD indexu odráží trend fluktuace dvou sousedních peak a trough bodů, toto je zachyceno v rovicích (3.7) a (3.8) [16].

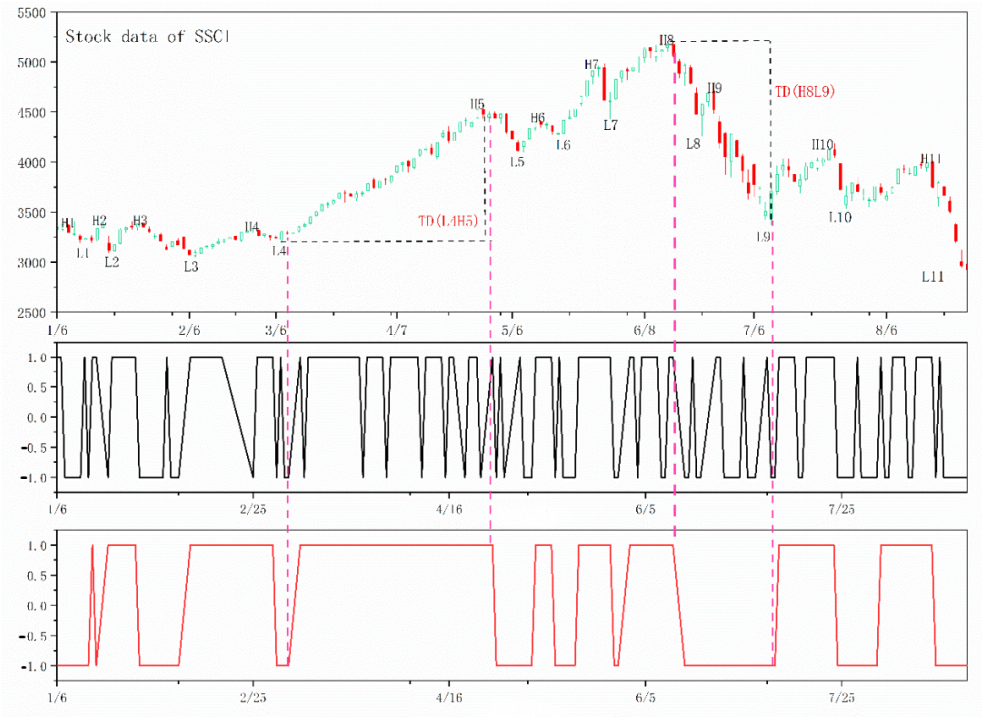
$$h = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_{t-1} \\ h_t \end{bmatrix} \qquad l = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_{t-1} \\ h_t \end{bmatrix} \tag{3.6}$$

$$TD(h_i l_{i-1}) = abs(\frac{h_i - l_{i-1}}{l_{i-1}}), i > 1 \tag{3.7}$$

$$TD(l_i h_{i-1}) = abs(\frac{l_i - h_{i-1} - 1}{h_{i-1}}), i > 1 \tag{3.8}$$

V algoritmu se porovnává volatilní parametr ω, který se porovnává s hodnotou TD. Kontinuální trend je v práci definován jako amplituda volatility dvou sousedních peak a trough bodů, překračující prahový parametr ω, jinak je volatilita považována za běžnou a bez kontinuálního trendu. Jako základ výpočtu jsou vzaty poslední nejnižší a nejvyšší hodnoty. Přesah ceny aktiva přes nebo pokles pod parametr ω je definováno jako kontinuální stoupající resp. klesající trend. Všem záznamům s rostoucím trendem je nastaven label na 1, zatímco klesajícím na hodnotu -1 [16].

Porovnání klasických labeling metod s metodou CTL je vidět v obrázku (3.3). Je zde vidět vypočtené TD časové řady a časové úseky L4H5 a H8L9. Na sekci L4H5 by šlo nahlížet jako na stálý kontinuální rostoucí trend. Avšak klasické metody obsahují mnoho šumu a tuto sekci rozdělí na několik sekcí rostoucích a klesajících trendů, stejně tak pro zbytek celé časové řady.



Obrázek 3.3: Zobrazení labelingu klasickou metodou a metodou CTL. První obrázek zobrazuje index SSCI. Druhý demonstruje labeling časové řady SSCI pomocí tradičních labeling metod. Poslední obrázek reprezentuje výsledek metody CTL. Časový úsek L4H5 je metodou označen jako dlouhodobě kontinuálně rostoucí trend, naopak H8L9 jako klesající trend. Z práce [16].

Každý investor má jiný názor na to jak vypadá kontinuální trend, dokonce i pro tu samou komoditu či akcii. Důvodem k odlišným pohledům jsou různé hodnoty jejich kapitálů, tolerance vůči risku, zkušenosti, rozdílné investiční strategie. Toto se snaží podchytit parametr ω. Pomocí jeho nastavování, mohou investoři provést data labeling na základě svých potřeb a natrénovat model, aby co nejvíce vyhovoval jejich potřebám a pomáhal s obchodováním. V následujícím pseudokódu je popsán algoritmus CTL z práce [16].

Algoritmus 1: Continuous trend labeling - Inicializace

Result: Časová řada s informacemi o kontinuálním trendu
Input : OHLC = [x1, ..., x_N]
$\omega > 0$ - parametr reprezentující poměr prahu, který definuje trend
Output: Labeling vektor y

$FP = x_1$; `// První cena`
$x_H = x_1$; `// Nejvyšší nalezená cena`
$x_L = x_1$; `// Nejnižší nalezená cena`
$TH = t_1$; `// Čas nalezení nejvyšší ceny`
$LH = t_1$; `// Čas nalezení nejnižší ceny`
$Cid = 0$; `// Identifikátor aktuálního směru trendu`
$FP_N = 0$; `// Index první nalezené nejvyšší/nejnižší ceny`
for $i = 1 : N$ **do**
  if $x_i > FP + x_1 * \omega$ **then**
    $[x_H, HT, FP_N, Cid] = [x_i, t_i, i, 1]$
    break
  end
  if $x_i < FP - x_1 * \omega$ **then**
    $[x_H, LT, FP_N, Cid] = [x_i, t_i, i, -1]$
    break
  end
end
$y_{0...FP_{N-1}} = Cid$;
$last\_label = FP_N$

Algoritmus 2: Continuous trend labeling - Labeling

Result: Časová řada s informacemi o kontinuálním trendu
Input : OHLC = [x1, ..., x_N]
$\omega > 0$ - parametr reprezentující poměr prahu, který definuje trend
Output: Labeling vektor y

```
for i = FP_N + 1 : N do
    if Cid > 0 then
        if x_i > x_H then
            [x_H, HT] = [x_i, t_i]
        end
        if x_i < x_H − x_H ∗ ω and LT ≤ HT then
            for j = 1 : N do
                if t_j > LT and t_j ≤ HT then
                    y_j = 1 last_label = j + 1
                end
            end
            [x_L, LT, Cid] = [x_i, t_i, −1]
        end
    end
    if Cid < 0 then
        if x_i < x_L then
            [x_L, LT] = [x_i, t_i]
        end
        if x_i > x_L + x_L ∗ ω and HT ≤ LT then
            for j = 1 : N do
                if t_j > HT and t_j ≤ LT then
                    y_j = −1 last_label = j + 1
                end
            end
            [x_H, HT, Cid] = [x_i, t_i, 1]
        end
    end
end
y_{last_label...N} = Cid;
```

3.3.4 Závěr

Jelikož jak již víme je trh přeplněn přeplněn algoritmickým tradingem a rychlé obchodování je velmi obtížné, protože stáhnout, zpracovat a vyhodnotit data je těžké oproti velkým hráčům, kteří dokáží data získat mnohem rychleji a se superpočítači je i rychleji vyhodnocovat, budu pracovat s metodou CTL, která se oproti prvním dvěma metodám zaměřuje na dlouhodobější trend.

3.4 Metriky

Pro vyhodnocení použiji matici záměn (3.3), která měří výkonnost klasifikačních modelů a metriky počítané z matice. Accuracy (3.9) – měří jak moc je klasifikátor přesný. Precision (3.10) – Když klasifikátor předpoví rostoucí trend, jak často je to správně. Recall (3.11) – Když je trend rostoucí, jak často klasifikátor předpoví rostoucí trend. F1 (3.12) – Vážený průměr mezi Recall a Precision.

Tabulka 3.3: Matice záměn. TP – správně klasifikovaný rostoucí trend, FP – špatně klasifikovaný rostoucí trend, FN – špatně klasifikovaný klesající trend, TN – správně klasifikovaný klesající trend

Skutečný trend / Predikce trendu	P (1)	N (−1)
P (1)	TP	FP
N (−1)	FN	TN

$$Accuracy = \frac{TP + TN}{P + N} \tag{3.9}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.10}$$

$$Recall = \frac{TP}{P} \tag{3.11}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.12}$$

Část II

Praktická část

Kapitola 4

Modelování a evaulace

Hlavním cílem praktické části je zpracovat a najít nejlepší formu dat pro další zpracování pomocí ML. Je zde brán důraz na transformaci Tick dat, data labeling, vyhodnocení pommocí ML modelu, pro různé metriky a v poslední řadě experimenty.

4.1 Technologie

Pro svou práci použiji skriptovací programovací jazyk Python verze 3. Ten je v poslední době nejpoužívanějším jazykem pro práci s daty [22], jejich analýzu a umělou inteligenci, což je přesně to čím se práce zabývá. Budu pracovat s velkými objemy dat, proto použiji knihovnou pandas, která je na toto připravena. Pro programování využiji Jupyter, což je webové vývojové prostředí, které umožňuje spouštět Python kód po částech v blocích, takže se nemusí celý vykonávat. Na ML budu využívat knihovnu scikit-learn (sklearn), která obsahuje širokou škálu klasifikačních, regresních a shlukovacích modelů pro Python.

4.2 Data

4.2.1 Zdroj

Dat je na burze obrovské množství. V této práci nepracuji s real-time daty, takže si vystačím s historickými daty. Ty jsou dostupné na internetu avšak velmi často jako již agregovaná data, nejčastěji na OHLC pomocí časového intervalu. Budu proto používat data z portálu *https://firstratedata.com/*, který obsahuje dataset tick dat amerických akcií, který je si třeba zakoupit. Jedná se o balíček *S&P 100 Tick Bundle*, který obsahuje data z let 2010-2021. Pro svou práci jsem si z důvodu velké náročnosti na strojový čas a práci s obrovskými daty vybral pouze část dat – 3 akcie AAPL, TSLA, MSFT a data za rok 2020 a 2021, kde se projeví i vliv COVID-19.

4.2.2 Popis a zpracování

Obsahují všechny obchody, jak v obchodních hodinách, tak i mimo ně. Všekerá data v datasetu jsou otestovaná a očištěná na odlehlé hodnoty (outlier), duplikátní a opožděné ticky. Každý soubor obsahuje o tick datech 4 atributy ve formátu (Timestamp (year-month-day hour:minute:second:millisecond), Price, Volume, Exchange), kde hodnoty jsou oddělené čárkami.

- Timestamp - datum a čas kdy obchod proběhl
- Price - cena za jakou obchod proběhl v dolarech
- Volume - počet akcií během obchodu
- Exchange - MIC (Market Identifier Code)

Poslední sloupec „Exchange“ je pro moje účely zbytečná informace, takže jej smažu, čímž ještě zmenším objem dat. Ukázka záznamů v následující tabulce (4.1).

Tabulka 4.1: Ukázka tick dat AAPL

Timestamp	Price	Volume	Exchange
2021-04-01 04:00:00:164	123.48	5	ARCX
2021-04-01 04:00:00:167	123.45	4	ARCX
2021-04-01 04:00:00:167	123.45	25	ARCX
2021-04-01 04:00:00:167	123.45	10	ARCX
2021-04-01 04:00:00:181	123.45	5	ARCX
2021-04-01 04:00:00:181	123.45	5	ARCX
2021-04-01 04:00:00:181	123.45	20	ARCX
2021-04-01 04:00:00:188	123.49	5	ARCX

Data jsou v jednotlivých .zip souborech po letech, přičemž každý soubor obsahuje 12 .csv souborů, které reprezentují jednotlivé měsíce v roku. V projektu jsem si vytvořil skript, který mi postupně jednotlivé roky rozbalil a udělal z každého .zip souboru jeden .csv soubor tím, že pospojoval jednotlivé měsíce. Takto vznikly soubory veliké až několik GB. Na velikosti souborů je vidět, že se na burze obchoduje čím dál tím víc. V následující tabulce (4.2) je vidět velikost Tick dat pro akcii Apple za během let.

Tabulka 4.2: Ukázka velikosti Tick dat během let akcie AAPL

2013	2016	2020	2021 (1Q)
750 MB	1.8 GB	5.1 GB	2.1 GB

Pro lepší představu vybraných parametrů pro transformaci Tick dat na Renko jsou v následující tabulce (4.3) uvedeny maximální a minimální ceny jednotlivých akcií v použitých datech.

Tabulka 4.3: Minimální a maximální ceny v datech jednotlivých akcií

	AAPL	MSFT	TSLA
Min	50.1	123.1	54.5
Max	153.4	270	906.4

4.3 Implementace

4.3.1 Modul stocks

Projekt obsahuje vlastnoručně implementovaný modul „stocks“, který zastřešuje veškerou důležitou funkcionalitu pro mou práci. Jsou zde implementovány následující funkce:

- time_bars(data, bar_minutes) – transformuje Tick data na OHLC v časovém intervalu - bar_minutes
- tick_bars(data, bar_ticks) – transformuje Tick data na OHLC podle počtu ticků - bar_ticks
- volume_bars(data, bar_volume) – transformuje Tick data na OHLC podle obchodovaného objemu - bar_volume
- dollar_bars(data, bar_dollars) – transformuje Tick data na OHLC podle množství obchodovaných dolarů - bar_dollars
- renko_bars(data, box_size) – transformuje Tick data na Renko podle změny ceny akcie - bar_size
- ctl(data, w) – continuous trend labeling metoda, která vrátí vektor s labelingem trendů v datech

4.3.2 Transformace Tick-by-Tick dat

Data všech 3 akcí jsou načtena a transformována pomocí metod z modulu „stocks“. Pro jejich transformaci jsou potřeba parametry na základě kterých se převedou na OHLC, každá akcie se obchoduje v jiném množství, za jinou cenu, v jiném objemu. . .Toto je potřeba vzít v úvahu. Všechny použité parametry pro transformaci Tick dat jednotlivých akcií jsou v následující tabulce (4.4). Jsou zvoleny tak aby vycházely počty svíček podobné u všech 3 akcií.

Tabulka 4.4: Parametry pro transformaci Tick dat na OHLC

Akcie	Bar minutes	Bar ticks	Bar volume	Bar dollars	Box size
AAPL	20	14000	3000000	200000000	4
TSLA	20	10000	1500000	300000000	15
MSFT	20	10000	1000000	200000000	4

4.3.3 Vyhodnocení

Na datech proběhl nejdříve data labeling metodou CTL, který označil v časových řadách klesající a rostoucí trendy. V tabulce (4.5) je vidět použitý parametr ω pro jednotlivé akcie, jelikož každá akcie má jinak se pohybující hodnoty, jsou použité rozdílné parametry.

Tabulka 4.5: Parametry CTL pro jednotlivé akcie

	AAPL	TSLA	MSFT
ω	5 %	8 %	5 %

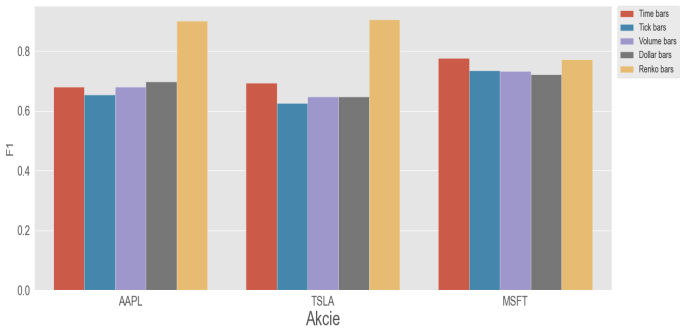
Poté byla rozdělena data na trénovací a testovací část pro kalsifikátor. Pro první část experimentu byl vybrán klasifikátor Random forest. Jedná se o složení více rozhodovacích stromů do jednoho klasifikátoru. Na tom byla natrénována data s trendy a následné použití pro predikci. K vyhodnocení vhodnosti svíček pro ML jsou v práci použity 2 faktory. Prvním faktorem je vhodnost pro předpovídání trendu, kterým se cena ubírá, což dává lepší pohled na to jak se cena hýbe a dát traderovi přehled o tom jak peníze investovat. K tomu jsou použity metriky z matice záměn (3.3).

V následující tabulce (4.6) jsou vidět metriky klasifikátoru, který byl natrénován na datech, aby předpovídal trend na jednotlivých typech svíček. Ze všech dopadly nejlépe Renko, které díky svému specifické metodě jejich vytváření pohlcují menší fluktuaci a dávají lepší přehled o trendu. Což je vidět i následujícím

grafu (4.1), který zobrazuje skóre F1 všech svíček, pro předpověď trendu všech 3 akcií.

Tabulka 4.6: Všechny metriky pro vyhodnocení různých typů svíček a předpovědi trendů na nich

Akcie	Typ svíčky	Accuracy	Precision	Recall	F1
AAPL	Time bars	0.646	0.674	0.688	0.681
	Tick bars	0.629	0.628	0.683	0.655
	Volume bars	0.650	0.659	0.702	0.680
	Dollar bars	0.650	0.682	0.713	0.697
	Renko	0.895	0.873	0.932	**0.901**
TSLA	Time bars	0.710	0.694	0.691	0.693
	Tick bars	0.667	0.652	0.599	0.625
	Volume bars	0.637	0.634	0.661	0.647
	Dollar bars	0.626	0.649	0.645	0.647
	Renko	0.891	0.899	0.912	**0.905**
MSFT	Time bars	0.726	0.762	0.790	**0.776**
	Tick bars	0.671	0.713	0.756	0.734
	Volume bars	0.680	0.724	0.742	0.733
	Dollar bars	0.676	0.728	0.715	0.722
	Renko	0.789	0.766	0.779	0.773



Obrázek 4.1: Porovnání skóre F1 předpovídání trendu pro jednotlivé akcie a typy svíček. Zdroj autor.

Druhým faktorem a experimentem, který rozhoduje vhodnost svíček pro

ML je, jak si vedl model při simulovaném tradingu na testovacích datech. Na datech byl metodou CTL proveden labeling, který určil trendy. Další metodou poté došlo k labelingu který značil signál – nákup (1) či prodej (−1), pomocí triviální strategie, kdy dojde k nákupu akcií, jakmile se změní trend z klesajícího na rostoucí a naopak při změně z rostoucího na klesající dojde k prodeji akcií. Došlo k rozdělení na train a test množinu, na kterých byla data se signály natrénována a následně vyhodnocena. Z důvodu značného množství dat, náročnosti na výpočetní výkon a strojového času nebyla použita křížová validace ke zlepšení výsledků modelu. Výsledky vyhodnocení jsou vidět v následující tabulce (4.7). Sloupec PPST značí – Poměr predikovaných a skutečných tradů.

Tabulka 4.7: Profity trading strategie simulované na predikovaných signálech RF

Akcie	Typ svíčky	Počet obchodů	PPST	Výnosnost (%)
AAPL	Time bars	2	0.4	9
	Tick bars	3	0.5	18
	Volume bars	5	0.55	51
	Dollar bars	6	0.85	61
	Renko	71	1	**135**
MSFT	Time bars	3	0.66	35
	Tick bars	3	0.66	25
	Volume bars	3	0.66	37
	Dollar bars	3	0.66	25
	Renko	94	0.8	**85**
TSLA	Time bars	8	0.7	120
	Tick bars	8	0.8	180
	Volume bars	15	0.85	**240**
	Dollar bars	9	0.8	150
	Renko	120	0.9	220

Pro další experiment byl vzat Multilayer perceptron klasifikátor, což je podtřída dopředných neuronových sítí, která se dokáže učit na nelineárních datech. Data byla pro klasifikátor standardizována. Výsledky tohoto experimentu v následující tabulce (4.8):

Tabulka 4.8: Profity trading strategie simulované na predikovaných signálech MLP

Akcie	Typ svíčky	Počet obchodů	PPST	Výnosnost (%)
AAPL	Time bars	4	0.8	33
	Tick bars	4	0.66	33
	Volume bars	5	0.55	52
	Dollar bars	5	0.7	45
	Renko	89	1	**135**
MSFT	Time bars	3	0.66	39
	Tick bars	3	0.66	43
	Volume bars	3	0.66	41
	Dollar bars	3	0.66	42
	Renko	94	0.8	**73**
TSLA	Time bars	8	0.13	-11
	Tick bars	8	0.62	95
	Volume bars	15	0.2	**98**
	Dollar bars	9	0.3	51
	Renko	113	0.7	70

4.3.3.1 Hyperparametry ML modelů

V této sekci jsou napsány hyperparametry modelů použitých v této práci.

Random forest klasifikátor

- bootstrap: True
- Max depth: 70
- Max features: auto
- N estimators: 2000

Multilayer perceptron klasifikátor

- α: 10^{-4}
- Neurony: 60
- Hidden layers: 7

4.4 Diskuze

Dle výsledků z práce je jasné, že Renko dopadlo celkově nejlépe. Jak na vyhodnocování trendu, kde se výsledky predikce trendu na svíčkách podobají výsledkům z práce [16], kde byla použita metoda CTL pro labeling, avšak výsledky nebyly zkoumány na různých typech svíček. Taktéž na predikci signálů na časové řadě, kde predikce na Renko dosahovala vyšších nebo podobných výsledků v porovnání s ostatními svíčkami.

Protože je v práci použita metoda CTL, která se přímo zaměřuje na trend, stejně tak Renko, je možné že při použití jiných metod data labelingu by výsledky mohly vyjít odlišně. Nebo při použitích jiných parametrů či trading strategií, které by používal trader s jiným smýšlením.

Machine learning, se učí z dat a snaží se v nich najít pattern. Je tedy otázkou, zda by výsledky vyšly stejně, kdyby v práci byla použita na učení i starší data než rok.

Kapitola 5

Nasazení

V této je popsán způsob nasazení a spuštění výsledného programu.

Jak již bylo v zmíněno, práce byla vypracovávána v Jupyter noteboocích. Stačí si tedy vytvořit virtuální prostředí, do kterého se nainstalují jednoduchým příkazem všechny potřebné knihovny. Jednou z nich je i Jupyter notebook. Ten lze následně spustit, což otevře webové rozhraní, ze kterého už lze spouštět notebooky a řídit veškerou práci. Všechny potřebné zdrojové kódy, data a podrobný návod na spuštění se nachází na přiloženém disku.

Závěr

Hlavním cílem této práce bylo zjistit, jaké časové řady, resp. jejich reprezentace svíčkami jsou nejvhodnější pro modely machine learningu pro generování buy/sell signálů. K tomu bylo důležité provést rešerši nejpoužívanějších metod převodu burzovních Tick dat na svíčky a data labeling metody na finančních časových řadách. V praktické části bylo třeba transformovat Tickova data, provést data labeling, následně modely natrénovat a na základě metrik vyhodnotit vhodnost svíček pro ML modely.

V první části práce jsem se zabýval umělou inteligencí resp. její podčástí machine learningem na finančních trzích, popsal jsem stručnou historii strojů na burze, výhody, aktuální stav a výhledy do budoucna.

V machine learningu je klíčové vybrat si kvalitní data pro daný účel, aby model byl co nejefektivnější. V rešerši jsem tedy popsal 5 různých dostupných reprezentací dat na burze a jak na tyto reprezentace převést očištěná tick data. Konkrétně se jednalo o časové, tickové, volume, dolarové a renko svíčky. Porovnal jsem jejich způsob vytváření, vlastnosti, výhody a nevýhody.

V další části rešerše jsem popsal 3 metody labelingu finančních časových řad pro machine learning, mezi které jsem zařadil naivní Fixed-time-horizon labeling sofistikovanější Triple barrier a nakonec CTL. Pro data labeling jsem vybral metodu CTL.

V poslední části rešerše jsou popsány metriky použité pro vyhodnocení vhodnosti svíček pro ML model.

V praktické části popisuji data se kterými pracuji a kde je získávám. Dále využité technologie a knihovny pro práci s daty, transformaci, labeling a ML modely. Jsou zde popsány parametry pro transformaci Tick dat na jednotlivé svíčky a parametry pro labeling metodu.

Vyhodnocení jednotlivých svíček na ML modelech ukázalo, že nejlépe modely pracují s renko svíčkami. Jak na předpovídání trendu, tak na simulovaném tradingu, kde ML model predikoval buy/sell signály. Naopak nejhůře dopadly časové svíčky, které mají nejhůře rozložené informace.

Téma práce je velmi rozsáhlé, kde já se zaměřil pouze na malou část

problému. Obsahuje tedy ještě mnoho místa pro další zkoumání a experimenty. Práce byla napsána tak, že s mírnou modifikací by mohla být použitá i na jiná aktiva než akcie. Do budoucna by šla rozšířit o analýzu svíček pro regresní problém předpovídání hodnoty. Přidání metody sliding window na časovou řadu.

Literatura

[1] Samuelsson. What Percentage Of Trading Is Algorithmic? online, březen 2021, [cit. 2021-03-27]. Available from: https://therobusttrader.com/what-percentage-of-trading-is-algorithmic/

[2] Kovacevic, A. The Evolution and Future of AI in the Stock Market. online, únor 2021, [cit. 2021-04-16]. Available from: https://hackernoon.com/the-evolution-and-future-of-ai-in-the-stock-market-nn2q33ou

[3] Swaine-Simon, S. The History of AI in Finance. online, březen 2018, [cit. 2021-04-16]. Available from: https://medium.com/district3/the-history-of-ai-in-finance-7a03fcb4a498

[4] Calcaterra, J. R. online, listopad 2020, [cit. 2021-04-14]. Available from: https://www.quora.com/Will-artificial-intelligence-end-the-stock-market-as-we-know-it

[5] Ülkü, N.; Onishchenko, O. Trading volume and prediction of stock return reversals: Conditioning on investor types' trading. online, únor 2019, doi:https://doi.org/10.1002/for.2582, [cit. 2021-03-20]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2582

[6] Working with High-Frequency Tick Data - Cleaning the Data. online, duben 2020, [cit. 2021-05-04]. Available from: https://quantpedia.com/working-with-high-frequency-tick-data-cleaning-the-data/

[7] High frequency data. online, září 2019, [cit. 2021-05-04]. Available from: https://en.wikipedia.org/wiki/High_frequency_data

[8] Venketas, W. Top 3 Technical Analysis Charts for Trading. online, duben 2019, [cit. 2021-03-12]. Available from: https://www.dailyfx.com/education/learn-technical-analysis/top-3-technical-analysis-charts-for-trading.html

[9] Kenton, W. Tick Definition. online, září 2020, [cit. 2021-02-20]. Available from: https://www.investopedia.com/terms/t/tick.asp

[10] Ivanov, M. Financial Machine Learning Part 0: Bars. online, únor 2019, [cit. 2021-02-16]. Available from: https://towardsdatascience.com/financial-machine-learning-part-0-bars-745897d4e4ba

[11] Mitchell, C. Tick Chart vs. One-Minute Chart: Which Is Better for Day Trading? online, červenec 2020, [cit. 2021-02-20]. Available from: https://www.thebalance.com/tick-chart-or-1-minute-chart-for-day-trading-1030978

[12] Folger, J. Advantages of Data-Based Intraday Charts. online, září 2020, [cit. 2021-03-05]. Available from: https://www.investopedia.com/articles/trading/10/data-based-intraday-chart-intervals.asp

[13] Reading Bar Charts for Trading: Time, Tick, Volume, Dollar, Information. online, květen 2020, [cit. 2021-03-12]. Available from: https://blog.quantinsti.com/bar-types-trading/

[14] Mitchell, C. Renko Chart Definition and Uses. online, listopad 2020, [cit. 2021-03-06]. Available from: https://www.investopedia.com/terms/r/renkochart.asp

[15] Hudgeon, D.; Nichol, R. Machine learning for business: using Amazon SageMaker and Jupyter. online, 2020, [cit. 2021-03-15]. Available from: https://aws.amazon.com/sagemaker/groundtruth/what-is-data-labeling/

[16] Wu, D.; Wang, X.; et al. A Labeling Method for Financial Time Series Prediction Based on Trends. online, říjen 2020, [cit. 2021-03-10].

[17] Ivanov, M. Financial Machine Learning Part 1: Labels. online, duben 2019, [cit. 2021-02-16]. Available from: https://towardsdatascience.com/financial-machine-learning-part-1-labels-7eeed050f32e

[18] Kenton, W. Beta. online, duben 2021, [cit. 2021-04-09]. Available from: https://www.investopedia.com/terms/b/beta.asp

[19] Gui, K. Data Labelling. online, říjen 2020, [cit. 2021-03-13]. Available from: https://towardsdatascience.com/the-triple-barrier-method-251268419dcd

[20] Wilcox, P. A Quantamental Approach Using Labeling for Stock Trading. online, březen 2019, [cit. 2021-03-14]. Available from: https://lucenaresearch.com/2019/03/27/quantamental-approach-to-stock-trading/

[21] Brown, S. J.; Goetzmann, W. N.; et al. The Dow Theory: William Peter Hamilton's Track Record Reconsidered. *The Journal of Finance*, volume 53, no. 4, prosinec 1998: pp. 1311–1333, doi: https://doi.org/10.1111/0022-1082.00054, [cit. 2021-04-05], `https://onlinelibrary.wiley.com/doi/pdf/10.1111/0022-1082.00054`. Available from: `https://onlinelibrary.wiley.com/doi/abs/10.1111/0022-1082.00054`

[22] Top 6 Data Science Programming Languages 2021 [Hand-Picked]. online, duben 2021, [cit. 2021-05-04]. Available from: `https://www.upgrad.com/blog/data-science-programming-languages/`

Kapitola 6

Zkratky

AI Umělá inteligence.

ATR Average True Range.

CTL Continous Trend Labeling.

ML Machine Learning.

MLP Multilayer perceptron.

OHLC Open-High-Low-Close.

RF Random forest.

Kapitola 7

Slovník

fluktuace Také volatilita, Míra kolísání hodnoty aktiva.

trend Trend je celkový směr ceny aktiva na trhu.

Seznam algoritmů

,

Kapitola 8

Obsah přiloženého CD

- `README.md` ..... Stručný popis obsahu CD
- `resources` ..... Tick data akcií AAPL, MSFT, TSLA
- `src` ..... Zdrojové kódy implementace
  - `notebooks` ..... Jupyter notebooky
  - `stocks` ..... Modul stocks se zdrojovými kódy
  - `thesis` ..... Zdrojová forma práce ve formátu LaTeX
- `text` ..... Text práce
  - `thesis.pdf` ..... Text práce ve formátu PDF