České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská

Katedra softwarového inženýrství
Obor: Aplikace softwarového inženýrství

Předvídání finančního trhu pomocí stromových metod

Financial market forecasting using tree based methods

BAKALÁŘSKÁ PRÁCE

Vypracoval: Pavel Ježek
Vedoucí práce: doc. Ing. Quang Van Tran, Ph.D.
Rok: Červenec 2022

České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská

Katedra softwarového inženýrství Akademický rok 2021/2022

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Student: **Pavel Ježek**

Studijní program: **Aplikace přírodních věd**

Obor: **Aplikace softwarového inženýrství**

Název práce česky: **Předvídání finančního trhu pomocí stromových metod**

Název práce anglicky: **Financial market forecasting using tree based methods**

Pokyny pro vypracování:

1. Charakterizovat strojové učení s důrazem na stromové metody.
2. Vypracovat rešerši na současný stav využití stromových metod k predikci finančních trhů.
3. Implementovat modely stromových metod a jejich využití pro predikci finančních trhů.
4. Vyhodnotit dosažené výsledky a srovnat je s výsledky podobných studií.

Doporučená literatura:

1) Kelleher, J. D., Mac Namee, B., D'arcy, A. (2020). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press. ISBN 978-0262044691.
2) Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* 4th Edition. San Francisco: Morgan Kaufmann. ISBN 978-0128042915.
3) Jiang, M., Liu, J., Zhang, L., Liu, C. (2020). An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms. *Physica A: Statistical Mechanics and its Applications*, *541*, 122272.
4) Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., Mosavi, A. (2020). Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. *IEEE Access*, *8*, 150199-150212.

Jméno a pracoviště vedoucího práce:

doc. Ing. Quang Van Tran, Ph.D.

Katedra softwarového inženýrství, Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

Jméno a pracoviště konzultanta:

.....

vedoucí práce

Datum zadání bakalářské práce: 1.10.2021

Termín odevzdání bakalářské práce: 7.7.2022

Doba platnosti zadání je dva roky od data zadání.

.....

garant oboru

.....

vedoucí katedry

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE Fakulta jaderná a fyzikálně inženýrská -1-

.....

děkan

V Praze dne 1.10.2021.....

Prohlášení

Prohlašuji, že jsem svou bakalářskou práci vypracoval samostatně a použil jsem pouze podklady (literaturu, projekty, SW atd.) uvedené v přiloženém seznamu.

V Praze dne ....

Pavel Ježek

Poděkování

Děkuji doc. Ing. Quang Van Tran, Ph.D. za vedení mé bakalářské práce a za podnětné návrhy, které ji obohatily.

Pavel Ježek

Název práce:
Předvídání finančního trhu pomocí stromových metod

Autor: Pavel Ježek

Obor: Aplikace softwarového inženýrství

Druh práce: BAKALÁŘSKÁ PRÁCE

Vedoucí práce: doc. Ing. Quang Van Tran, Ph.D.
Katedra softwarového inženýrství, Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

Abstrakt: Práce se zabývá predikcí na finančním trhu pomocí stromových metod. Zkoumá umělou inteligenci, strojové učení a jeho dělení. Dále také je přiblížena struktura stromu, stromové metody, jejich přesnost a souborové metody k ním přidružené. Jelikož se jedná o rostoucí odvětví, je zde připojena řešerše stávajících prací na téma predikce pomocí strojového učení. Vytvořili jsme několik stromových modelů a náhodných lesů z dat indexů FTSE a S&P pro období 2010-2019. Jejich přesnosti byly následně porovnány a s nejlepšími bylo provedeno modelové obchodování.

Klíčová slova: Predikce finančího trhu, S&P, FTSE, stromové metody

Title:
Financial market forecasting using tree based methods

Author: Pavel Ježek

Abstract: The thesis deals with prediction in the financial market using tree methods. It explores artificial intelligence, machine learning and its subdivisions. The structure of the tree, tree methods, their precision and the file methods associated with them are also approximated. As this is a growing industry, a survey of existing works on the subject of prediction using machine learning is attached. We created several tree models and random forests from FTSE and S&P index data for the period 2010-2019. Their accuracies were then compared and model trading was performed with the best ones.

Key words: Financial market prediction, S&P, FTSE, tree methods

Obsah

Úvod

Strojové učení je rychle rostoucím odvětvím, hlavně v posledních letech. Často se skloňuje s termínem umělé inteligence, ale má mnoho využití. Mezi ty se řadí například: výcvik samořídících robotických jednotek, vytváření rozpoznávacích algoritmů (rozpoznávání obrazu, zvuku atp.), nebo také predikce. A na ty se zaměřuje tato práce, přesněji na predikce finančního trhu pomocí stromových metod a jejich nástavby v podobě souborových metod složených z nich (náhodných lesů). Strojové učení dokáže nalézt, z dat, která jsou mu dána, různé spojitosti mezi atributy na finančních trzích. Může tedy nahradit některé z různých ekonomický analýz, které bývají náročné na čas.

V 2.kapitole jsou popsány základní pojmy, kterými jsou umělá inteligence a strojové učení. Obsahuje definice jednotlivých pojmů, jakou spojitost mezi sebou mají, jakou historií si tento pojem prošel nebo co za cíl si tato disciplína dává. U strojového učení je zde popsáno i dělení na kategorie.

Ve 3.kapitole jsou detailněji blíže prozkoumány jednotlivé kategorie strojového učení. Rozšiřují tyto pojmy, co znamenají a pro jaké druhy problémů jsou koncipovány. Dále také jak se modely rozdělují na další podkategorie. Dále je zde například popsáno, jak by měla být vstupní data upravena nebo zdali vůbec dané modely to vyžadují.

4. kapitola se zabývá stromovými metodami, které jsou tématem této práce. Vysvětluje se zde co je to strom a jak vypadá jeho struktura. Opět se zde popisují jednotlivé druhy stromů a následně také základní tvorba stromu. To zahrnuje rozdílné přístupy na pravidla větvení problému ve struktuře. Dále jsou zde přiblíženy způsoby, jak počítat přesnost či chybu a jak zlepšit strom. Na to navazuje i popis souborových metod složených z jednotlivých stromů.

Současnému využití strojové učení pro predikce finančního trhu se zabývá 5. kapitola. V té se nachází rešerše vědeckých prací na toto téma. Ty jsou kategorizovány podle použitých metod, popřípadě vstupních dat. Také se zde objevuje krátké vysvětlení ostatních analýz trhu, jakou je např. fundamentální.

6. kapitola již popisuje samotné vytvořené modely. Nejdříve se uvádějí použité nástroje a data. U dat se přibližuje i v jakém formátu jsou zpracovávána. Následují výsledky jednotlivých modelů. To zahrnuje jejich přesnosti, jakými optimální hyperparametry byly vytvořeny či jakým způsobem byly nalezeny. Poslední části této kapitoly je interpretace výsledků, která je ukázána na příkladu obchodování v části testovacích dat.

Kapitola 1

Umělá inteligence a strojové učení

1.1 Definice umělé inteligence

Myšlenka tvorby umělé inteligence už lidstvo provází již desetiletí. I když je těžké určit přesné datum, začátky této myšlenky se dají vystopovat do roku 1952, kdy americký spisovatel Isaac Asimov publikoval v časopise Astounding Science Fiction povídku Hra na honěnou a s ním 3 zákony. Tyto nápady se časem přeměnily ze žánru sci-fi do reality a inspirovaly vědce jak v oborech umělé inteligence, tak i robotiky nebo počítačových věd [1].

I dnes se však na téma definice umělé inteligence, nebo inteligence samotné, vedou rozsáhlé diskuze. Například Oxfordský slovník definuje umělou inteligenci jako: *„The capacity of computers or other machines to exhibit or simulate intelligent behaviour; the field of study concerned with this. Abbreviated AI.“* [2]. V překladu tato definice znamená: *„Schopnost počítačů nebo jiných strojů vykazovat nebo simulovat inteligentní chování; studijní obor, který se toho týká. Zkráceně AI.“*. Díky velkým pokrokům v tomto odvětví se často definice mění. Většina definicí se dá zobecnit do 4 kategorií: systémy myslící jako lidé, jednající jako lidé, myslící racionálně a jednající racionálně [3].

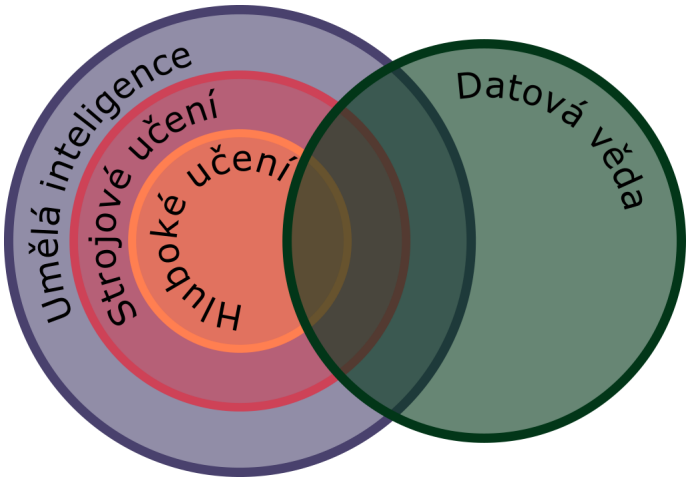
Když tedy je definován pojem umělé inteligence, nastává problém, jak zjistit že daný systém lze nazvat umělou inteligencí. Ten řeší například Standartní Turingův test. Autor Allan Turing ho publikoval v roce 1950 a prezentoval ho na tzv. Imitační hře. Hra byla popsána jako jednoduchá hra pro 3 hráče. Máme hráče A a hráče B, kde jeden z nich representuje ženu a druhý muže. Třetí hráč zastupuje roli rozhodčího a může být jakéhokoliv pohlaví. Hráč C snaží pomocí otázek zjistit pohlaví ostatních hráčů s omezením na komunikaci pouze písemnou. Zároveň tázaní hráči mají určenou roli, kdy hráč A se snaží zmást hráče C a hráč B mu pomoct. Turing poté vyslovil otázku: *„Co se stane, když roli A v této hře převezme stroj? Rozhodne se vyšetřovatel špatně stejně často, když se hraje takto, jako když se hraje mezi mužem a ženou?"Tyto otázky nahrazují naši původní, „Dokážou stroje myslet?„‘*. Společně s touto hrou publikoval v roce 1950 ještě jednu obdobnou s rozdílem, že hráči A i B mají za úkol se snažit mystifikovat hráče C [4].

Tyto testy však nejsou neomylné a bez kritiky. Jednou výtku uvedla Susan G. Sterrett, že není přesně určeno, která z výše uvedených formulací hry je ta, kterou Turing zamýšlel test přemýšlení počítače. Mezi nevýhody těchto testů můžeme zařadit například problém, že ne všechna rozhodnutí člověka jsou inteligentní, a zároveň racionální chování nemusí být považováno jako lidské rozhodnutí [4].

1.2 Definice strojového učení

Jak už z názvu vyplývá, definice strojového učení bude souviset s definicí učení, jako takového. Sloveso „učit se"je definováno jako souhrn několika vlastností. Jedná se o: získávání znalostí o něčem studiem, zkušenostmi nebo výukou, zapamatovat si přijaté znalosti, získat povědomí informacemi nebo pozorováním, být informovaný nebo snažit se zjistit informace, schopnost přijímat instrukce. Tato definice však narazí na problémy, pokud se jí budeme snažit použít pro počítač. Problém je například nalézt způsob, jak zjisti, že se systém něco naučil anebo získal povědomý povědomí pozorování. Další otázkou je způsob určení uvědomění své existence počítače, to už ale zasahuje více do filozofické roviny. Tuto definici lze však přeformulovat, do více strojově přívětivější podoby. Potom význam učení subjektu je, když mění svoje chování tak aby fungoval (nebo pracoval) lépe v budoucnu. Tím pádem se jako sledovaným faktorem jeví účinnost. Zde je možné už vidět spojitost, proč dané odvětví získalo svůj název [5].

Strojové učení je studování algoritmů, které se samy vylepšují na základě zkušeností nebo přijatých dat. Je viděno jako podtřída umělé inteligence. Na základě typu přijímaných zkušeností se jednotlivé algoritmy se dělí na 3 hlavní kategorie: učení s učitelem a zpětnovazebné učení. Mezi tyto se zavádí ještě další, tzv. učení bez učitele, kombinace učení s učitelem a bez učitele. Jedná se o kombinaci prvních dvou kategorií. Hlavní cílem strojového učení je zobecňovat problém na základě přijatých dat a snažit se o co nejpřesnější předpověď při poskytnutí dříve neviděných dat [6].



Obrázek 1.1: Grafické znázornění vztahů mezi Umělou inteligencí, Strojovým učením, Hlubokým učením a Datovou vědou

Kapitola 2

Dělení strojového učení

2.1 Učení s učitelem

Učení s učitelem (angl. „Supervised learning“) je kategorie algoritmů strojového učení, kde tréninková data jsou ve tvaru *D(x; y)*. V proměnné x je uložen zadaný dataset a v proměnné y je uložené označení které koresponduje s danými daty (máme n dat, potom existuje n dvojic $(x_1; y_1), \ldots, (x_n; y_n)$ kde k datům x_i je přiřazeno označení y_i, kde $i \in (1, \ldots, n)$. Znamená tedy že data jsou nějakým způsobem kategorizovány, příkladem může být trénování programu pro rozpoznávání objektů před kamerou, kde za x jsou uloženy různé informace o fotografii objektu a za y je název objektu.

Základem těchto algoritmů je hledání funkce $g : X \to Y$, kde X je množina dat a Y je množina označení (nebo názvů). Jedná se o funkci, která k novému data setu x vybere nejvhodnější označení y. Funkce g je součástí tzv. prostoru hypotéz G. Někdy je výhodně funkci g vyjádřit pomocí hodnotící funkce $f : X \times Y \to R$, potom g je definována jako $g(x) = argmax_y\ f(x,y)$. Kde f je prostor hodnotících funkcí. Avšak mnoho algoritmů strojového učení jsou pravděpodobnostní modely, kde $g(x) = P(y|x)$ a $f(x,y) = P(x,y)$. K volbě optimální funkce g se používají 2 základní funkce minimalizace rizika. Jde o Empirickou minimalizaci rizika a Strukturální minimalizaci rizika. První z nich hledá funkci nejvíce odpovídající datům, a druhá používá tzv. trestní funkci, která vytváří kompromis mezi zkreslením a rozptylem.

Tento typ učení se dělí ještě na dvě další subkategorie, a to na klasifikaci a regresi [6].

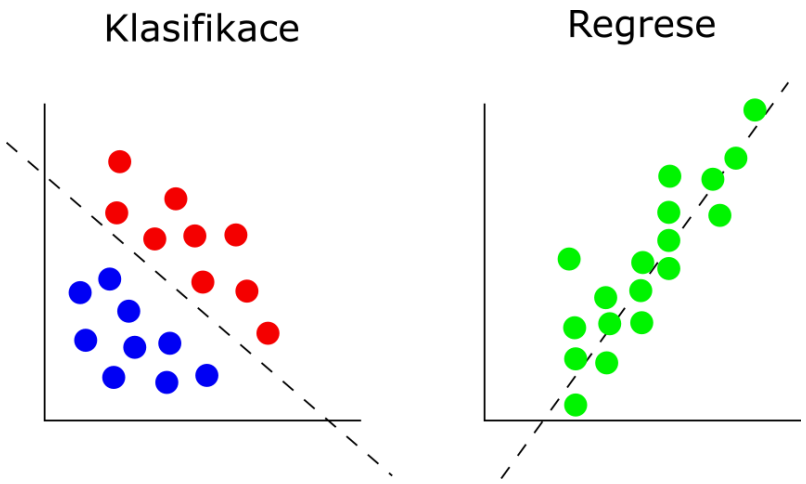
2.1.1 Klasifikace

Klasifikační algoritmy se snaží přiřadit dat x k označení y. Množina označení má tvar $\{1, \ldots, n\}$, kdy čísla znamenají různá označení (kategorie). Pokud $n = 2$, jedná se o binární klasifikaci, pro $n > 2$ se používá označení vícetřídní klasifikace. Některé příklady používají i více označení pro jednu část data setu (např. na obrázku je vyfocena osoba, která je vysoká a zároveň mladého věku – dvě kategorie),

v tomto případě mluvíme o více značkové klasifikaci. Do této podkategorie spadají jak lineární, tak nelineární modely. Příkladem algoritmů jsou algoritmus Naivní Bayes, Metoda podpůrných vektorů nebo Stromové metody. Klasifikační algoritmy jsou nejvíce využívanou formu strojového učení [7].

2.1.2 Regrese

Regresní algoritmy se podobojí klasifikačním, pouze s rozdílem, že y je vyjádřeno spojitě. Přesněji jde o snahu najít vztah mezi závislými (cílovými) a nezávislými (odpovědnými za předpověď) proměnnými, jak změny v závislé ovlivňují nezávislou (pokud jich je více, dívá se na jednu z nich a zbytek zůstává neměnný). Při regresi tedy vytvoříme graf z dat a snažíme se o nejlepší fit funkcí skrz daná data neboli: *„Regrese ukazuje čáru nebo křivku, která prochází všemi datovými body na grafu cíle-předpověď takovým způsobem, že vertikální vzdálenost mezi datovými body a regresní přímkou je minimální"* [8]. Mezi ty algoritmy patří například Lineární, Logistická nebo Polynomiální regrese [9].



Obrázek 2.1: Grafické znázornění typů dat užitých pro klasifikaci a regresi

2.2 Učení bez učitele

Algoritmy učení bez učitele zpracovává data, která nebyla předem jakkoliv označena. Proto algoritmy sami hledají různě se vyskytující vzory a korelace mezi jednotlivými daty. Rozdílem tedy je že těmito modely se snažíme předpovídat více proměnných než jenom jednu u učení s učitelem. Tyto algoritmy jsou také nejblíže k pojetí učení u lidí a zvířat. Má také větší obecné využití než učení s učitelem, protože nepotřebuje předem kategorizovat data označením. Jsou také lépe využitelná na komplexní problémy, protože nám řeknou více dat, než je to u prvního druhu učení. Opět se dělí na dvě podkategorie: asociace a shlukování.

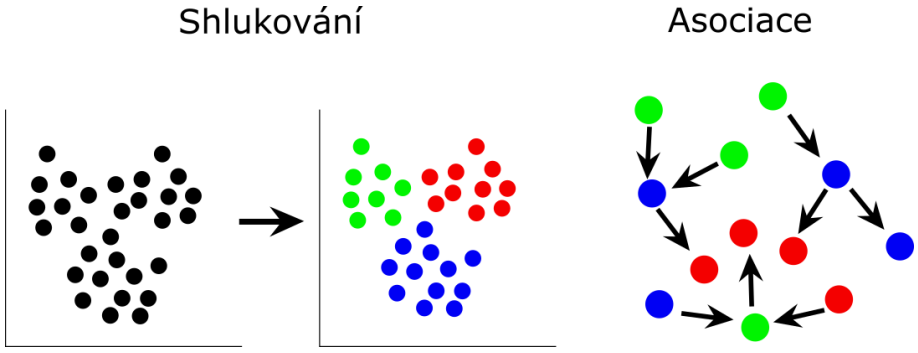
2.2.1 Asociace

Jedná se o metody zaměřené na asociativní pravidla. Snaží se o nalezení různých korelací a spojení mezi daty ve velkých databázích. Prochází různé kombinace jednotlivých dat se vytvářejí pravidla a ta se vybírají na základě jejich pokrytí (počet správně předpovězených případů) a přesnosti (počet případu ku celkovému počtu). Tyto různé kombinaci vytvářejí tzv. sety položek, z těch pomocí přednastavených minimálních hodnotách pokrytí a přesnosti vybíráme pravidla, ty se utvářejí různými kombinacemi položek v setu položek. [5].

2.2.2 Shlukování

Tyto algoritmy shromažďují objekty do shluků (clusters), tak že mají více společných rysů než objekty z ostatních skupin. Základním modelem je shlukování podle vzdálenosti. V tomto případě se pomocí jednotlivých částí dat vypočítá vzdálenost mezi jednotlivými objekty (nejčastěji se používá euklidovská vzdálenost). Shlukování se dělí do dalších 4 podkategorií: exklusivní, překrývající, hierarchické a pravděpodobnostní shlukování.

V prvním typu se jednotlivé objekty zařadí vždy do jednoho shluku a nemůže být členem jiného. V druhém naopak objekty mohou být členy více shluků a díky tomu se tyto skupiny překrývají. Objekty v tomto typu mají u odpovídající míru zavázání k shlukům, ve kterých jsou členem. Třetí typ využívá spojování sousedních shluků, kdy jsou udány parametry ukončení sjednocování. Na začátku těchto algoritmů jsou všechny body určeny jako shluky. Poslední typ používá pravděpodobnostní metody k určení shluků. [10].



Obrázek 2.2: Grafické znázornění shlukování a asociace, šipky v části asociační vyjadřují různé spojení mezi daty

2.3 Kombinace učení s učitelem a bez učitele

Tato kombinace používá data, která jsou z části označená (stejně jako u učení s učitelem) tak i data u kterých není předem dáno označení, proto spadá to meziprostoru

mezi prvními dvěma kategoriemi. Tento přístup kombinace označených a neoznačených dat může vyprodukovat poměrně přesné výsledky, pozitivem je také úspora na práci při označování dat. Tyto algoritmy jsou pak více efektivní, než kdyby se použily pouze označená data a učení s učitelem nebo neoznačená data a učení bez učitele. Dělí se na dva typy: transduktivní a induktivní učení. První hledá označení pro neoznačená data, a druhé hledá správné určení jaká data patří do jakého označení (hledá správné zobrazení z množiny dat do množiny označení).

Aby se dali využít neoznačená data, musí se pro ně použít jeden ze tří předpokladů. Prvním je předpoklad kontinuity, to se spoléhá na to, že velice podobné body mají stejnou značku (body nacházející se blízko sebe – geometricky vyjádřeno). Dává se přednost hranicím jednotlivých tříd označení v méně hustě obsazené části data setů, proto je málo bodů s jinými štítky blízko sebe. Druhý je předpoklad že data se stejným štítkem shlukují. Třetím je mnohostranný předpoklad, kde se většina dat nachází v menší množině (prostoru) něž je vstupní množina(prostor). Příkladem může být rozpoznávání hlasu jednoho člověka, kdy zvuk jako takový má velký počet možných frekvencí, ale hlas oné osoby se pohybuje pouze v nějakém rozmezí [11].

2.3.1 Zpětnovazebné učení

Hlavní myšlenkou těchto algoritmu vytvoření sekvenci rozhodnutí. Určitý trénovaný systém je vystaven komplexnímu prostředí a v něm řeší problém podobající se hře. Počítač využívá metody pokusu a omylu, aby vyřešil danou překážku. Pro upřesnění cíle se využívá systém trestů a odměn, avšak systém nedostává žádné návody, jak danou situaci projít. Počítač nedostává žádná označená data, proto se může učit pouze zkušenostmi. Cílem systému je vylepšovat efektivitu a tím získat největší odměnu [13].

Kapitola 3

Stromové metody

Stromové metody nebo také metody rozhodovací stromů spadají pod učení s učitelem, jak je možné vidět výše. Nejčastěji je spojováno s klasifikačními algoritmy, ale používají se i regresní stromy. Rozhodovací stromy jsou uspořádané struktury, které využívají strategii rozděl a panuj. Ta je definována jako, hledání optimálního řešení pomocí rozdělení problému na 2 a více podproblémů. Tento postup se rekurzivně opakuje, dokud rozdělením nevznikne jednoduše splnitelný požadavek. Obecné se dají rozdělit na jednorozměrné a vícerozměrné stromy [14].

3.1 Základní struktura

Jak už bylo zmíněno jedná se o hierarchickou strukturu. Jedná se o sekvenci rekurzivních rozdělení problému na menší části. Ta jsou vytvořena pomocí rozhodnutí, které se ve stromové struktuře nacházejí v uzlech, poslední uzel se nazývá koncový nebo listový. Každý uzel m obsahuje testovací funkci $f_m(x)$ s diskrétně označenými potomky. Po vstupu informace do uzlu se pomocí rozhodovací funkce postoupí do odpovídajícího potomka. Proces začíná u kořene stromu a končí u listové vrstvy, kde informace označená v končícím udává výstup algoritmu. Stromové metody se považují za neparametrické, jelikož se neudává žádná podmínka na složitost dat nebo velikost stromu [14].

3.2 Dělení stromů

3.2.1 Jednorozměrné stromy

Jednorozměrné stromy mají v každém uzlu pouze testy atributů jedné dimenze. Pokud se rozděluje pomocí diskrétního algoritmu, rozhodnutí se bude větvit na n cest. U numerických (seřazených) atributů se používá binární rozdělení (máme číslo x a rozdělujeme data na menší nebo větší než x, tedy rozdělení na intervaly). Pokud u nominálního atributu rozdělení pokrývá všechny možnosti tohoto atributu, nazý-

váme ho kompletní. Nejrozšířenějšími jsou klasifikační a regresní stromy. Hlavními rozdíly jsou mezi nimi styly dat, kde u klasifikačních stromů máme diskrétní a u regresních spojité. S tím je spojeno i rozdílné stavění stromů [14][15].

3.2.2 Vícerozměrné stromy

V uzlech těchto stromů, má testovací funkce více dimenzí, a dokonce může použít všechny možné. Pokud je vstup numerický, můžeme použít lineární binární vícedimenzionální uzel. Ten je definován jako:

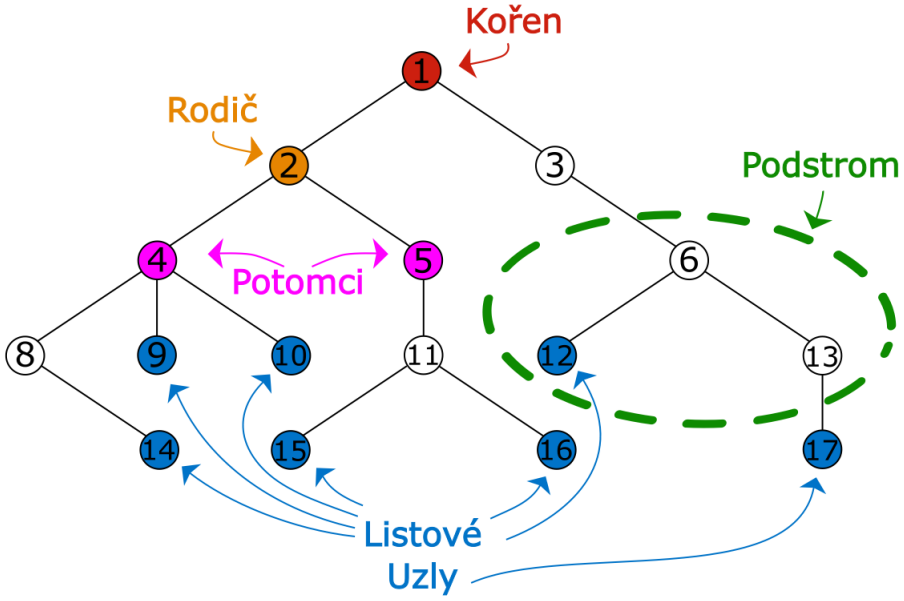
$$f_m(x) : w_m^t x + w_{m0} > 0 \tag{3.1}$$

Kde w je hmotnostní vektor a w_0 je prahová hodnota. Uzly v jednorozměrném stromu jsou tedy speciální variantou (kde všechny kromě jednoho atributu, jsou nuly).

Dalšími přístupy jsou například kvadratické uzly (podobné lineárním pouze s přídavkem kvadratické veličiny) nebo sférické uzly. Druhé jmenované jsou definovány pomocí:

$$f_m(x) : \|x - c_m\| > \alpha_m \tag{3.2}$$

Kde c_m a c_m je střed a α_m je poloměr. Algoritmů používající tento přístup je více, příkladem může být vícerozměrná verze algoritmu CART (Classification and regression trees) [14].



Obrázek 3.1: Základní struktura stromu

3.3 Tvorba stromu

Jak vyplývá ze strategie rozděl a panuj, strom se skládá z různých rozdělení problému. A z toho vychází i problém, jak vybrat atribut, podle kterého se bude dělit na podproblémy. Metod existuje více a každá se využívá na jiný typ problému.

3.3.1 Klasifikační stromy

U klasifikačních stromů se hledá čistota (purity) jednotlivých uzlů. Tuto metriku nazýváme Informace a jednotky jsou bity. Vyjadřuje, jak jednotná je třída odpovědí v jednotlivých potomcích (jestli se do potomka uzlu dostanou pouze data s jedním typem označení a v druhém případě, jak je tam zastoupena ta nejpočetnější). Udává nám množství informace potřebné k rozhodnutí, zda nová instance je jedním nebo dalším objektem z množiny označení. Většinou se tyto hodnoty pohybují pod 1 bitem, na rozdíl od paměti, kde se využívá stejná jednotka [5].

Informace a Informační zisk

Jelikož daná veličina potřebuje, aby splňovala následující podmínky:

- Pokud ve zkoumané uzlu je pouze zastoupen jeden druh označení, potom je informace nulová
- Pokud mají všechny druhy označení stejné zastoupení, informace hodnota informace maximální.

Tyto podmínky splňuje funkce s názvem Informační entropie:

$$entropy\,(p_1,\ p_2,\ \ldots,\ p_n) = \ -\ p_1 \log\,(p_1) - p_2 log(p_2) - \ldots - p_n log(p_n) \qquad (3.3)$$

Argumenty $p_1,\ \ldots,\ p_n$ jsou zlomky, kde jmenovatelem je počet dat patřící k jedné značce a čitatelem je celkový počet dat proudících do daného uzlu (jejich součet tedy musí být číslo 1).

Nejčastěji se používá logaritmus se základem 2. V případě výpočtu informace se používá značení:

$$info\,([a_1,\ \ldots,\ a_n]) = entropy\left(\frac{a_1}{m},\ \ldots, \frac{a_n}{m}\right) = entropy(p_1, \ldots,\ p_n) \qquad (3.4)$$

kde $a_1,\ \ldots,\ a_n$ jsou počty zastoupení jednotlivých instancí se stejným označením a m je celkový počet instancí proudících do uzlu.

To nám však řekne informační hodnotu pouze potomků, ale pro správný výběr budeme potřebovat vědět, jak přínosný je daný uzel, podle kterého budeme rozdělovat.

To nám řekne informační zisk. Pro jeho výpočet potřebujeme zjistit průměrnou hodnotu všech potomků:

$$info\left([a_1,\ \ldots,\ a_{k1}],\ \ldots,\ [r_1,\ \ldots, r_{kr}]\right) = \left(\frac{\sum a}{n}\right) * info\left([a_1,\ \ldots,\ a_{k1}]\right) + \ldots\ + \left(\frac{\sum r}{n}\right) * info\left([r_1,\ \ldots, r_{kr}]\right) \quad (3.5)$$

Kde n je celkový počet instancí, r je počet potomků a $a_1, \ldots, a_{k1}$ jsou počty instancí jednoho druhu označení v potomkovi A (příp. $r_1, \ldots, r_{kr}$ jsou počty instancí jednoho druhu označení v potomkovi R).

Informační zisk uzlu y s potomky a až r vypočteme pomocí:

$$gain(y) = info\left([x_1,\ \ldots,\ x_k]\right) - info\left([a_1,\ \ldots,\ a_{k1}],\ \ldots,\ [r_1,\ \ldots, r_{kr}]\right) \quad (3.6)$$

Kde y je jeden z atributů data setu (ten se nachází v rodičovském uzlu potomků $A, \ldots, R$), $x_1, \ldots, x_k$ jsou všechny druhy označení a_i pro $i \in 1, \ldots, k_1$ (příp. r_j kde $j \in (1, ..., k_r)$ jsou počty instancí jednoho druhu označení v potomkovi A (popř. potomka R) [5].

Poměr zisku

Tento přístup má však svoje negativa, kde zvýhodňuje atributy, které zastupuje velké množství odlišných hodnot. Příkladem může být například identifikační číslo, kde každá instance má vlastní číslo. Potom při tomto atributu se tedy bude algoritmus nejvíc větvit, proto se využívá úprava pomocí výpočtu poměru zisku.

Spojením těchto výpočtu získáme Poměr zisku (IGR):

$$IGR(y) = \frac{gain(y)}{info([\sum a,\ \ldots,\ \sum r])} \quad (3.7)$$

Kde *gain(y)* udává informační zisk uzly y a $info([\sum a,\ \ldots,\ \sum r]))$ nám říká vnitřní informační hodnotu uzlu s atributem y (tedy jeho potomků).

I tato modifikace může vést pro špatný výběr, proto se k tomu porovnává informační zisk daného atributu k průměru informačních zisků ostatních zkoumaných rozhodujících atributů [5][12].

Gini index

Dalším častým způsobem rozhodovací hodnoty je Gini index. Jedná se opět měření čistoty, respektive v tomto případě pravděpodobnost nesprávného zařazení, tedy spíše „nečistoty“. Hodnota se pohybuje mezi 0 a 1. Hodnota 0 je nabývána, pokud je v listovém uzlu pouze jeden druh (kategorie) atributu a 1 když jsou zde zastoupeny všechny. Hodnota 0,5 nám udává, že hodnoty jsou rovnoměrně rozloženy do

některých kategorii. Index se vypočítá jako:

$$Gini = 1 - \sum_{k=1}^{n} p_k^2 \tag{3.8}$$

Kde p_k pravděpodobnost výstupu (kategorii) k v uzlu. Při rozdělení se Gini index počítá pro každého potomka. Celková hodnota Gini indexu je potom rovna váženému součtu Gini indexu v potomcích.

$$Gini = \sum_{i=1}^{m} \frac{n_i}{n} Gini(i) \tag{3.9}$$

Kde m je počet potomků, n_i je počet pozorování v i-tém uzlu, n je počet pozorování v rodičovském uzlu a *Gini(i)* udává Gini index v i-tém uzlu [19].

3.3.2 Regresní stromy

Stavba regresních stromů se podobá té u klasifikačních, avšak kritériem není čistota, ale rozptyl (snaha o jeho snížení).

Střední kvadratická chyba

Označíme X_m množinu všech atributů X, které projdou od kořene do uzlu m. Definujeme funkci $b_m(x)$, která nabývá hodnot 1 a 0 (1 pro $x \in X_m$ a jinak 0). Jako hlavní metrika rozdělení se využívá střední kvadratická odchylka:

$$E_m = \frac{1}{N} \sum_t \left(r^t - g_m\right)^2 b_m(x^t) \tag{3.10}$$

Kde g_m odhadovaná hodnota uzlu, N je celkový počet hodnot v X_m, r je požadovaný výstup (neboli označení). Odhadovanou hodnotu uzlu určíme jako:

$$g_m = \frac{\sum_t b_m(x^t) r^t}{\sum_t b_m(x^t)} \tag{3.11}$$

Pokud odchylka E_m po výpočtu zůstává pod zadanou konstantou (neboli je míra variance uzlu v přijatelné hodnotě), potom je vytvořen listový uzel a je do něj uložena hodnota g_m. Avšak je-li hodnota vyšší, dojde k tvoření dalších podstromů. Snažíme se dělit tak, aby součet chyb potomků byl minimální. Podle tohoto pravidla rekurzivně hledáme atributy (resp. u numerických hodnot intervaly nebo nějaký bod zlomu), pro ty opět počítáme jak E_m tak i g_m, ale už pouze pro část X_m (jelikož potomci si původní množinu dat X_m rozdělí na více menších, což vyplívá z definice rozděl a panuj).

Nejhorší možná chyba

Pro rozdělení můžeme také použít například nejhorší možnou chybu:

$$E_m = \max_j \max_t |r^t - g_{mj}| b_{mj}(x^t) \tag{3.12}$$

Kde index m_j označuje potomka původního uzlu m, r požadovaný výstup a g_m odhadovaná hodnota uzlu. Tato hodnota nám udává že žádná další instance není větší než dané omezení [14].

3.4 Vylepšování stromu a jeho přesnost

3.4.1 Prořezávání a validace

Sestavené stromové struktury někdy obsahují nepotřebné části, které nemají zásadní vliv na. Těch se dá zbavit pomocí tzv. prořezávání (pruning). Tento přístup můžeme použít jak při tvorbě stromu (prepruning), nebo až po jeho sestavení (postpruning) [5].

Prepruning

Pokud algoritmus využívá prepruning, většinou tvoří podmínku na základě minimálního počtu instancí. Pokud do tohoto podstromu vstupuje méně než daný počet instancí, podstrom se nevznikne. Tento přístup, však nepere v potaz, že může kombinace více atributů mít zásadní informační, ale samotné atributy nemají při rozhodování velkou váhu. Dále neřeší, že algoritmus nezkoumá již vytvořené uzly (tzv. backtracking). Nejčastěji se využívá, pokud je klíčovým faktorem výpočetní složitost [5].

Postpruning

Tato metoda se zpracovává až po sestavení stromu, jedná se tedy o více procesně náročnou operaci. Ale díky tomu má lepší výsledky než prepruning. V tomto postupu se využívají dvě různé metody: Nahrazení podstromu a zvýšení podstromu.

Primárně využívané je nahrazení podstromu. Hlavní myšlenkou je nahrazení „špatného“ podstromu uzlem. Prochází stromem od listové vrstvy až po kořen. Druhým způsobem je zvýšení podstromu. Jedná se o více komplexní přístup a není jasné, zda vždy bude mít větší přínos. Jak už z názvu vyplývá, jedná se o zvýšení potomka na místo rodiče (rodičovský uzel zanikne a místo něj se napojí potomek). Samozřejmě, zbývající potomci rodičovského uzlu se musí zahrnout nového uzlu. Jedná se o velice náročnou operaci, proto se neprovádí na složitých uzlech.

Tento algoritmus musí zjistit, který z uzlů je potřeba upravit a jakou modifikaci použít. Jedním z přístupů může být využití části tréninkových dat, které se vyčlení

před sestavením stromu. Pomocí nich se pak kontrolují jednotlivé uzly [5].

Kritérium složitosti

Jak vyplívá z uvedených metod, velikost stromu souvisí s jeho složitostí. Zároveň s rostoucí velikostí může strom ztrácet obecnou platnost, ale příliš malý zase naopak může pokrývat pouze malé množství případů. Určujeme tedy kritérium složitosti:

$$C_\alpha(T_1) = DT_1 + \alpha |T_1| \tag{3.13}$$

Kde T_1 je prořezaný strom (T_0 je původní strom), DT_1 je jeho chyba, $/T_1/$ je počet terminálních uzlů a α je parametr pro kompromis mezi přesností a velikostí stromu. Cílem je tedy minimalizace tohoto kritéria, tedy vhodné α. To se určuje pomocí křížové validace [19].

Křížová validace

Jedná se o metoda, která zjišťuje, jak moc nezávislé vzorky dat ovlivňují model. Data rozdělíme na k nezávislých částí. Vždy se sestaví model z *k-1* podsouborů a zbývající se použije na testování. Tím máme vytvoříme k stromů testovaných na jiných datech. Z těchto modelů můžeme zjistit např. průměr a směrodatnou odchylku nebo její schopnost předpovědět. Jelikož jsou stromové struktury nestabilní, změny v datech mohou způsobit velké změny ve stromové struktuře i její přesnosti. Z těchto modelů vybereme ten s největší přesností a s nejmenším rozdílem chyby testovacích a tréninkových dat, nám určí hodnotu alfa [19].

3.4.2 Přesnost

U křížové validace je snaha o zjištění o co nejlepšího stromu pro určení koeficientu alfa, a proto je potřeba vypočítat přesnost stromu. Výpočet této hodnoty se liší podle typu stromu [19].

Klasifikační stromy

Jedna z využívanějších a jednodušších metod pro výpočet přesnosti u klasifikačních stromů je celková správnost:

$$OA = \frac{n_p}{n} \tag{3.14}$$

Kde n_p jsou správně klasifikovaná data a n je celkový počet. To však není úplně dostačující, protože nezohledňuje velikost skupin nebo rozdíly oproti náhodnému výsledku. Pak mohou vycházet přehnané výsledky. Příkladem je, když jedné výstupní kategorie je násobně více než dalších, potom nám pro tuto kategorii mohou

vyjít vysoké hodnoty přesnosti, navzdory tomu že ostatní výstupní atributy strom určuje velice špatně. Využívá se tedy korekce na tyto velikosti:

$$OA = \frac{1}{K} \sum_{c=1}^{K} \frac{n_{pc}}{n_c} \tag{3.15}$$

Kde K je počet výstupních kategorií, n_{pc} jsou správně klasifikovaná data pro výstup c, n_c je celkový počet dat pro výstup c. Celková chyba se potom využívá, spolu s dalšími klasifikačními metody, pro určení optimálního stromu.

Regresní stromy

Pro určení přesnosti se u regresních stromů využívá koeficient determinace *R2*. Obecně je tento koeficient využíván u lineární regrese a vypočítá se jako podíl variability závislé proměnné k celkové variabilitě modelu. U stromových metod se tedy jedná o podíl kvadratických odchylek predikce ku kvadratickým odchylkám průměru, které jsou odečteny od 1.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{n} (y_i - \overline{y_i})^2} \tag{3.16}$$

Kde $\widehat{y_i}$ je průměr v terminálních uzlech (predikovaných hodnot), $\overline{y_i}$ je průměr v kořenových uzlech, a y_i je k tomu adekvátní data (pozorování). Hodnota se pohybuje mezi 1 a 0, pokud se rovná 1 potom predikce se shodují s daty.

Lze vypočítat chybu regresního stromu. Pro tréninkovou sadu:

$$e(t) = 1 - R^2_{tren} \tag{3.17}$$

A pro testovací platí:

$$e(t) = 1 - R^2_{test} \tag{3.18}$$

Kde $R^2{}_{tren}$ a $R^2{}_{test}$ jsou koeficienty determinace pro tréninkovou a testovací soubor.

3.4.3 Chyba

Pro pozorování, jaký ze stromu je ten optimální využíváme i chybu měření. Pro tu existuje optimistický a pesimistický odhad. Optimistickým odhad pro terminální uzel je myšleno:

$$e(t) = e(t) \tag{3.19}$$

a pesimistickým odhadem je:

$$e(t) = e(t) + 0,5 \tag{3.20}$$

Kde *e(t)* je chyba na tréninkovém souboru a *e´(t)* je chyba na testovacím souboru. Máme tedy dva odhady chyby stromu. Ty se spojí a vytvoří se celková chyba:

$$e(t) = e(t) + 0,5 * N \tag{3.21}$$

Kde *e´(t)* a *e(t)* jsou chyby na testovacím a tréninkovém souboru, N je počet terminálních uzlů [19].

3.5 Souborové metody

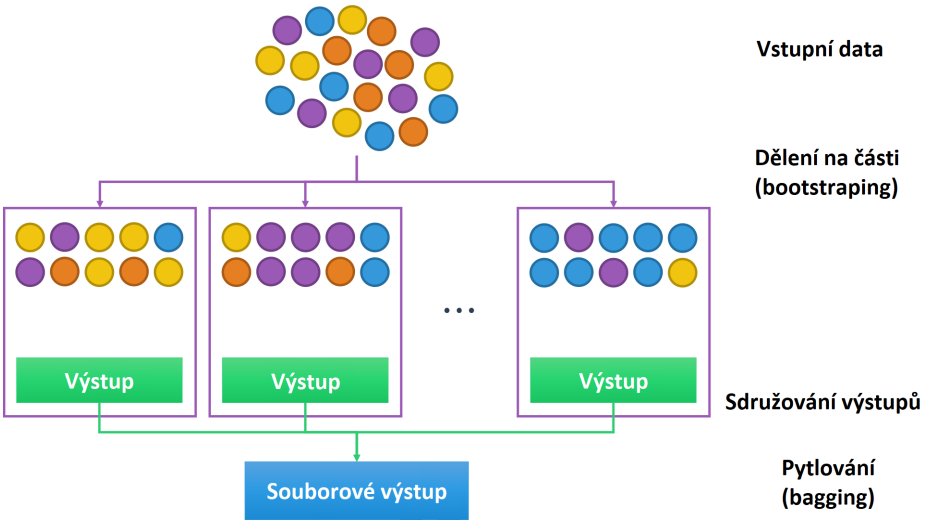
Souborové metody vychází z Condorseho pravidla poroty. Originálně se vztahoval pouze na dva možné případy výstupu, ale v oboru strojového učení je myšlenka rozšířena na více. Základní principem tohoto pravidla je že rozhodování více subjekty, i když méně spolehlivějšími, docílíme větší přesnosti. V strojovém učení se tedy problém zpracuje více algoritmy (např. více rozhodovacími stromy), proto se jim říká i učení závislé na porotě nebo vícenásobné klasifikační systémy. Tyto méně přesné modely se označují jako slabí žáci, a jsou to většinou právě modely vytvořené klasickými metodami strojového učení (označují se také jako základní). Metody tedy dokážou vylepšit předpovědi z těchto slabých žáku na předpovědi silných žáků (modely s lepší přesností predikce).

Mezi nejčastěji používané patří agregace bootstrapu, posilování (boosting) a náhodný les (random forest) [21][22].

3.5.1 Agregace bootstrapu

Jedná se o souborový meta-algoritmus. Využívá se hlavně u rozhodovacích metod, ale lze ji využít i jako samostatný přístup. Základní premisou je rozdělení modelu na několik menších.

Máme soubor tréninkových dat D velikosti n. Ten rozdělíme na m části o velikosti n. Každá z nich může obsahovat pouze část z původních dat (cca 63,2 %), ale data se mohou (resp. musí) opakovat. Tyto části se nazývají bootstrapové výběry. Těchto m bootstrapových výběrů se zpracuje (Agreguje – sdruží) a jejich výstup se zhodnotí (hodnotí se většina nebo průměr) [16] [22].



Obrázek 3.2: Grafické znázornění rozdělování, přeloženo [17].

3.5.2 Posilování

Posilovací metody (boosting) se zaměřují na snížení zkreslení a rozptylu. Základní principem je iterativně tvořit stromy, kde datům jsou přiřazovány váhy. Ty se odvíjejí podle toho, jak si vedly v předchozí iteraci (jak špatně byla zařazena, čím častěji tím jsou důležitější). Výstupem potom jsou tedy predikce:

$$\widehat{y}(x) = \sum_{t} w_t h_t(x) \tag{3.22}$$

Kde $h_t(x)$ je výstup stromu t a w_t je váha daného atributu. Cílem je minimalizovat funkci:

$$O(x) = \sum_{i} l\left(\widehat{y}_i, y_i\right) + \sum_{t} \Omega(f_t) \tag{3.23}$$

Kde $\widehat{y}_i$ je předpověď a y_i je pravdivá hodnota, $(\widehat{y}_i,\ y_i)$ je vzdálenost mezi hodnotami a $\Omega(f_t)$ je regulační funkce (penalizuje složitost stromu) [22] [23].

3.5.3 Náhodný les

Náhodný les (random forest) je rozšířením agregace bootstrapu. Využívá se pro klasifikační i regresní operace. Hlavním principem je tvorba několika rozhodovacích stromů a práce s jejich výstupy (u klasifikačních se rozhoduje jaké řešení má nejvíce hlasů, u regresních zase průměr hodnot).

Tvorba náhodného lesa

Jak už bylo řečeno výše, jedná se a nadstavbu na agregaci bootstrapových výběrů. Teda data jsou rozděleny na m bootstrapových výběrů. Rozdílem je, že stromy vy-

bírají prvky náhodně, a proto nedochází k přehnané závislosti rozhodnutí na tréninkových datech (tzv. overfiting). Z těchto m souborů se vytvoří stromy. Přitom proces zkoumá, jak počet (nebo absence) prvků v bootstrapovém výběru ovlivňuje výsledek. Tyto informace pak využívá ve výpočtu tzv. matice záměn (confusion matrix, dále česky chybová matice). Díky nim pak klasifikujeme jednotlivé atributy pomocí různých metrik (např. informační zisk popsaný výše). Postup s porovnáváním atributů se rekurzivně opakuje pro další úrovně stromů. Výstupní hodnota je pozitivní, pokud byl pozitivní i poslední zkoumaný a negativní pro opak. Tvorbu lesa můžeme vylepšit či zrychlit různými hyperparametry (parametry zadávána před průběhem trénování). Ty můžou ovlivňovat velkou škálu vlastností. Může jím být minimální počet parametrů v listové vrstvě nebo maximální počet atributů, při kterých dojde k rozdělení [16].

Chyba předpovědi

Pro výpočet chyby předpovědi se využívá například metoda Oob (Out-of-bag). Při této metodě se vyrábějí další výběry, k již vytvořeným bootstrapovým. Ty zahrnují atributy, které nebyly vloženy do bootstrapových výběrů. Po vytvoření stromů se data z Oob souborů nechají projít těmito stromy a hodnotí se, jak přesně se předpoví výsledek.

To se pak dále využívá pro měření důležitosti jednotlivých atributů. Kdy vypočítáme Oob chybu pro každý prvek a zprůměrujeme ho přes celý les. Pak se prvek různě prohází do tréninkového data setu a opět se počítá Oob. Změny v přesnosti ve stromech po permutaci prvku se zprůměrují. Tato hodnota nám pak říká, jak významný je prvek na rozhodovacím procesu (misclasification rate, značení MR). Často se vyjadřuje v procentech, kdy 100 % má nejdůležitější prvek. Pokud je tedy prvek bezvýznamný pro rozhodnutí, bude mít nízkou hodnotu MR.

Dalším indikátorem může bát i změna na predikci, kdy pozorujeme, jak se změnilo rozložení ve výstupech mezi jednotlivými stromy. Tedy počet stromů, který predikovali stejnou hodnotu (kategorii).

$$zp_{cv} = cpv - \ cpv_{Ran} \tag{3.24}$$

Kde zp_{cv} je změna v pravděpodobnosti, cpv je pravděpodobnost jednotlivých kategorii (jestli máme 100 stromu a z toho 15 má stejný výstup, potom pro danou kategorii platí $cpv = (15/100)$ a cpv_{Ran} je pravděpodobnost kategorii pro jednotlivou permutovanou proměnou. Tento postup provedený pro všechny výstupy nám pak může ukázat různé korelace mezi atributy a výstupy. [18] [19].

Těsnost

Náhodné lesy se dají využít i pro shlukování (clustering), k němu se využívá metrika těsnosti (proximity). Po sestavení stromů se pozoruje listová vrstva, a hledáme které atributy se zde vyskytují spolu. Pokud se pozorované atributy vyskytnou ve stejném koncovém uzlu, vzroste jejich těsnost. Hodnota těsnosti je potom udávána jako počet

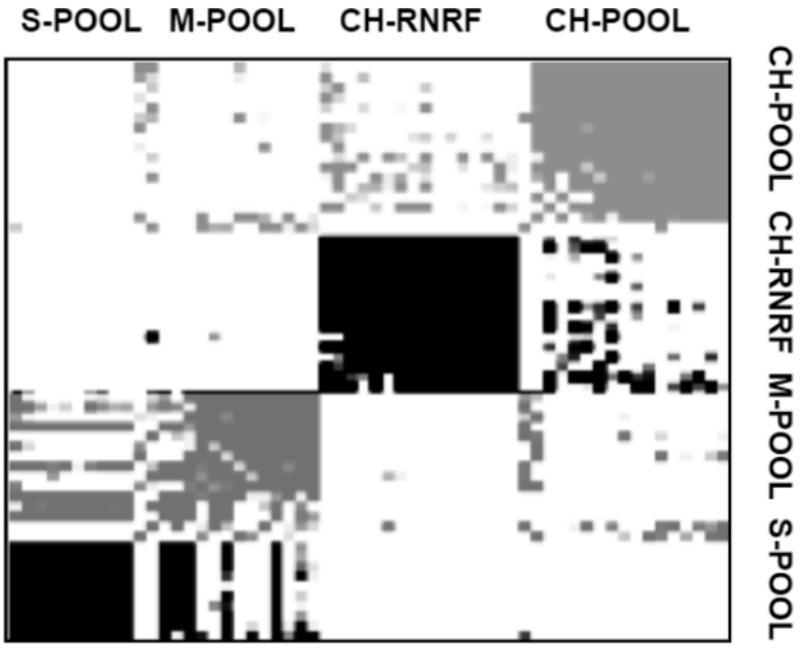
výskytu podělený celkovým počtem stromů (tedy pokud se objeví pár spolu v 1 z 10 stromů je hodnota 1/10). Z toho vyplývá že hodnota nemůže být větší než 1, přičemž 1 značí maximální těsnost, tedy pár se objevil spolu v listovém uzlu ve všech stromech.

Tuto těsnost využíváme například na zobrazení překrývání kategorií, které pak zobrazujeme v teplotních mapách (probability heat map). K vytvoření mapy lze využít i matici pravděpodobností. Dalším využitím může být měření odlehlosti výstupního atributu od ostatních. To se vypočítá jako:

$$out(n) = \frac{N_j}{P(n)} \tag{3.25}$$

Kde N_j je počet pozorování j-tého výstupního atributu a $P(n)$ je průměrná těsnost pozorování j-tého atributu. Pokud je hodnota větší jak 10 jsou brány je jako odlehlé.

Těsnost se taky využívá u nahrazování chybějících hodnot. Hodnota se nahradí váženým průměrem ostatních hodnot stejného typu, a jako váhá se využívá míra těsnosti mezi nahrazovaným prvkem a hodnot využívaných k jeho doplnění. Tyto hodnoty jsou využity v dalších stromech. Tento postup se opakuje, dokud lze řešení vylepšit nebo může být ohraničeno pevnou hodnotou. Pro nahrazování se může použít i pouze medián hodnot, ale ten má menší přesnost než postup využívající těsností [19].



Obrázek 3.3: Příklad teplotní mapy, čím tmavší barva tím větší hodnota těsnosti[19]
.

Kapitola 4

Současné využití strojového učení pro predikce

Strojové učení je dnes hojně využíváno v mnoha odvětvích. Využívají i velké mezinárodní společnosti, Kterým pomáhají s nakládáním s velkým množstvím dat. Ty po zpracování mohou využít na mnoho funkcí, které museli dříve zastupovat lidé. Příkladem využívání je rozpoznávání obrázků (využívaný společností Meta v jejich úspěšné sociální síti Facebook pro označování přátel na fotografiích), rozpoznávání řeči (domácí asistenti, kteří dokážou vykonávat příkazy vyslovené uživatelem), autonomní ovládání robotů (samořídící automobily, vyvíjené například firmou Tesla) nebo například různé předpovědi. Ty se zaměřují na různá odvětví, ať už se jedná o marketing, meteorologie nebo trh s akciemi [24].

Pro předpovědi na trhu s akciemi se může využívat více druhů metod. Jendou z nich je Fundamentální analýza. Ta hledá rozdíly mezi tržní cenou a „pravou“ cenou. Pro tento postup využívá makroekonomické i mikroekonomické faktory (Tržby, prodeje, ...). Nepozoruje pouze faktory týkající se pouze předpovídaného subjektu, ale i celkový stav trhu a odvětví ve kterém se subjekt nachází. Nalezená „pravá“ cena je pak porovnávána s tržní a na základě toho se zjišťuje, zda je firma nadhodnocena nebo podhodnocena trhem. Další metodou je například technická analýza. Ta předpokládá, že data o minulém chování subjektu na trhu mohou pomoc s předpovědí jejího budoucího chování. Více se tedy zaměřuje na množství a cenu akcií [25][26]. S rozvojem výpočetní se začal využívat i další způsob zahrnující strojové učení. Na to se zaměřuje rozpor vědeckých článků.

Práce a vědecké články jsou zaměřeny na predikce akciovém trhu pomocí strojového učení. Jsou z období posledních jedenácti let. Uvedené články (vědecké práce) byli čerpány ze ScienceDirect, IEEE a Google Scholar. Ty byly rozděleny podle využívané metody, nebo typu vstupních dat. Uvedené články nejsou reprezentací celého odvětví, ale jedná se o průřez tím nejvyužívanějším.

4.1 Umělé neuronové sítě

Umělé neuronové sítě (Artificial neural networks, zkráceně ANN) jsou velice využívanou metodou pro predikce finančního trhu. Při této metodě se vytváří model snažící se napodobit lidský mozek. Pro problém je vytvořena struktura s označenými uzly (neurony) spojenými váženými cestami (neuronové propojení). Data se tímto systémem opakovaně, kdy při každé nové iteraci se předá i chybovost. Tou se model vylepšuje, dokud nedojde požadované přesnosti [27].

První vybraná práce od trojice autorů Erkam Guresena, Gulgun Kayakutlua a Tugrul U.Daimb vydaná v roce 2011. Ta se zaměřuje na použití několika druhů metod neuronových sítí. Jedná se o vícevrstvý percepton (MLP), dynamický umělé neuronové sítě (DAN2) a na selekci vstupních dat hybridní neuronové sítě využívající zobecněnou auto regresivní podmíněnou heteroskedascticitu (GARCH). S použitím Střední kvadratické odchylky a střední absolutní odchylky pak vypočítají chyby jednotlivých modelů. Pro predikci si vybrali NASDAQ index v časovém období 2008-2009. Zjistili, že pro jejich daný porovnávací set, bylo nejlepší obyčejné MLP bez použití GARCH [28]. Další práce, vypracovaná Bing Yangem, Hao Jiankunem a Zhang Sichang v roce 2012, se zabývá použitím modelů neuronových sítí pro předpovědi na Šanghajském akciovém trhu. Ve své práci využily model zpětné propagace neuronových sítí. Ty využívá skrytých vrstev uzlů, které nejsou spojeny se vstupem a výstupem, pro minimalizaci chyb. Vstupními daty byly denní ceny Šanghajského složeného indexu burzy měsíce v roce 2010. Jejich práce ukázala že Šanghajský akciový trh je dobře predikovatelný daným modelem, avšak to přikládali tomu, že čínský trh s cennými papíry není efektivní [29]. O předpovědi trhu pomocí umělých neuronových sítí napsali také autoři Amin Hedayati Moghaddama Moein, Hedayati Moghaddam a Morteza Esfandyari. Byla vydána v roce 2016. Tato studie měla za úkol schopnosti umělé neuronové sítě pro predikce denního směnného kurzu NASDAQ z roku 2015, a jaký rozdíl je mezi přesností při změně úpravy dat pro různé počty předcházejících dnů [30].

4.2 Metoda podpůrných vektorů

Druhou hojně využívanou metodou pro predikce je metoda podpůrných vektorů (zkr. SVM). Hlavní myšlenkou tohoto postupu je vytvoření prostoru prvků s vysokou dimenzí ze vstupních dat (vektorů). V něm se pak vytvoří rozhodovací plocha. Její vlastnosti pak pomáhají zobecnit daný problém [31].

Tento způsob predikce využily ve své práci Shunrong Shen, Haomiao Jiang a Tongda Zhang ze Stanfordské univerzity z roku 2012. Jako vstupní data vybrali informace z několika burz v období 2000-2012, kde hlavními byly NASDAQ, DJIA a S&P. Výsledné hodnoty pro tyto hlavní tři indexy se pohybovaly v číslech vyšších než 74 %. Dále porovnaly modely vytvořené podle metody podpůrných vektorů se základními nákupními strategiemi na burze [32]. Další práci s touto metodou napsali v roce 2013 Linkai Luo a Xi Chen. Zaměřuje se na integrování lineární reprezentace a metody vážených podpůrných vektorů pro predikci signálu na burzovním trhu. Li-

neární reprezentaci využily na zjištění hlavních bodů změny a jakou váhu tyto změny měli na budoucí průběh. Poté se použije metoda podpůrných vektorů s určenými váhami. Ten potom vyhodnotí model, ke kterému pak autoři přidali i indikátory pocitů investorů pro vylepšení přesnosti předpovědi. Vstupní daty bylo 20 typů akcií z Šanghajské burzy v letech 2005-2006. Své výsledky porovnaly s metodou využívající neuronové sítě a metodou „buy and hold“, kde jejich přístup byl vyhodnocen jako nepřesnější a stojí za další prozkoumání [33]. Uplatnění této metody pro predikce popsali také Pan Yuchen, Xiao Zhi, Wang Xianning a Yang Daoli. Práce byla vydána v roce 2017. V ní se zabývali problematikou více výstupů a navrhly nový způsob přístupu s podporou více výstupů a neomezeným vzorkováním smíšených dat. Byla to jedna z prvních prací, které tyto kombinace postupů užívá [34].

4.3 Další metody a jejich kombinace

I Když jsou neuronové sítě a podpůrné vektory oblíbenými postupy pro predikce pomocí strojového učení, mají svoje nevýhody. Nejsou tedy univerzálně nejlepšími metodami, a proto se využívají i jiné. Avšak všechny přístupy mají svoje nevýhody, proto se také propojuje několik metod dohromady.

4.3.1 Jiné metody

První z této kategorie je práce z roku 2017 od trojice autorů Eunsuk Chong, Chulwoo Han a Frank C. Park. Ta byla zaměřena na predikce na korejském akciovém trhu pomocí hlubokých neuronových sítí (podtyp umělých neuronových sítí). Jako vstup použily data z let 2010-2012, s pětiminutovou frekvencí (Intra denní informace). S porovnáním s umělými neuronovými sítěmi byly výsledky lepší, ale ne ve všech případech (za rozdíly byly různé postupy etrakce parametrů. Většinou se rozdíly však vyrovnaly při užití testovacího souboru. Autoři poukázali na fakt, že pro hluboké učení jsou vhodnější více přesná data (s vyšší frekvencí). Hluboké neuronové sítě, podle autorů, mají využití v oblasti predikce trhu, ale mají své nevýhody a je potřeba její další výzkum [35]. S hlubokými neuronovými sítěmi také pracovali i Xiao Zhong a David Enke v roce 2019. Vytvořily analytický proces předpovídající denní vývoj indexu S&P. Jako vstupní data využily 60 různých atributů (náležící S&P) z období 2003-2013. Predikce vytvářeli pro různé hodnoty skrytých vrstev a ty pak porovnaly s hodnotami s obecnými umělými neuronovými sítěmi. Dospěli k závěru, že použití hlubokých neuronových sítích je přesnější než umělé neuronové sítě, ale poukázali i na jejich větší výpočetní náročnost z důvodu vyššího počtu neuronů (uzlů) s přibývajícími skrytými vrstvami [36].

Neuronové sítě byly použity i ve vědecké práci od Jigar Patel, Sahil Shah, Priyank Thakkar a K. Kotecha. V té použily nejen neuronové sítě (přesněji umělé neuronové sítě) ale také metodu podpurných vektorů, náhodný les a Bayesovu naivní metodu. Všechny 4 následně porovnaly. Predikce byly vytvořeny na základě dat z období 2003-2012, jednalo se o akcie dvou firem, index CNX Nifty a index S&P bombajské směnné burzy. Zaměřením tohoto dokumentu bylo i na rozdíl jaká data se předpoví-

dají. První přístup byl zaměřen na predikci 10 technických parametrů ekonomických subjektů a druhý na reprezentaci budoucího trendu těchto subjektu. Pro první zmíněný postup byl vyhodnocen náhodný les jako nejpřesnější s více jak 83 % přesností a u druhého Bayesova naivní metoda s více jak 90 % přesností (u druhé zmiňované došlo ke zlepšení u všech algoritmů) [37]. Další uplatnění strojového učení popisují i M. Umer Ghania, M. Awaisa a Muhammad Muzammul. Tuto studii vydali v roce 2019. Jsou zde užívány 3 postupy: Lineární regrese, metoda tříměsíčního pohyblivého průměru a exponenciální vyhlazování. Vstupní data jsou akcie Amazonu a Applu z první poloviny roku 2019. Jejich výsledné predikce byly pak porovnány a nejlepších výsledků dosáhla metoda exponenciálního vyhlazování [38].

Všechny výše zmíněné odborné práce využívaly pouze data z různých akciových trhů, ale se stále větší vlivem médií a sociálních se pro zpřesnění výsledků využívá i analýza těchto vlivů – analýza sentimentu. Tyto dvě předpovědi využily Mazhar Javed Awan, Mohd Shafry Mohd Rahim, Haitham Nobanee, Ashna Munawar, Awais Yasin a Azlan Mohd Zain v článku z roku 2020. Práce předpovídala 10 akcií známých firem pomocí lineární regrese, zobecněné regrese, rozhodovacích stromů a náhodného lesa. Dvě z nich využila na analýzu sentimentu, která byla tvořena pomocí logistické regrese a Bayesova naivního modelu. Po propočtení modelů se přesnost lineární regrese, zobecněné regrese a náhodného lesa pohybovala kolem 90 % (s výjimkami), zatímco rozhodovací stromy se ukázali jako neefektivní a překročily hranici 80 % pouze u jedné z akcií. Sentimentální analýza (které předpovídá, zdali bude akcie stoupat či klesat a napomáhá tedy investorů v rozhodování) se pohybovala v rozmezí mezi 70-80 %, kde nejvyšších hodnot nabývala Bayesova naivní metoda [39]. Další práce, s podobným námětem, integruje sentimentální analýzu přímo do předpovědi. Napsali ji Wasiat Khan, Mustansar Ali Ghazanfar, Muhammad Awais Azam, Amin Karami, Khaled H. Alyoubi a Ahmed S. Alfakeeh. Vstupními daty bylo 11 různých ekonomických subjektů (indexy burz i jednotlivé akcie firem) z různých burz, k tomu byly přidány data ze sociálních sítí a finančních médií. Předpověď potom byla provedena na 12 různých modelech. Jednalo se o Gaousovského-Bayesovu naivní metodu, mnohočlenné Bayesovo naivní metody, metodu podpůrných vektorů, logistické regrese, vícevrstvý perceptron, k-nejbližší sousedi, klasifikační a regresní stromy, lineární diskriminační analýzu, metodu AdaBoost, Klasifikátor zesílení přechodu, náhodný les a metodu Extra tree. Po vytvoření a úpravě modelů (pomocí redukce spam zpráv a redukce nepotřebných atributů dat) je nejlepším (a nejvíce konzistentním) náhodný les. Dále práce vyhodnotila že ne všechny ekonomické subjekty jsou stejně ovlivňovány sociálními médii a finančními zprávami, nebo že některé jsou více volatilní, a tudíž hůře předvídatelné (např. akcie MSFT byly shledány jako velice proměnlivé a média na ní mají velký vliv) [40].

4.3.2 Kombinace

Jak už bylo řečeno výše, z důvodu omezení negativních vlastností jednotlivých metod se využívají jejich kombinace. V současnosti je často brána kombinace společně s genetickým algoritmem. Ten vychází z evoluční biologie (Evoluční teorie – Charles Darwin). Algoritmus vytváří různá řešení, které pak procházejí přes speciální

funkci (ta je analogií k přežití nejsilnějšího v přírodě), která je ohodnotí. Subjekty s nejsilnějšími relativními hodnotami jsou pak „kříženy“. Vytváří se potomek dvou vybraných subjektů (stejně jako v biologii, jedná se o spojení části informace z obou předků). Aby se předcházelo hledání pouze v oblasti lokálního optima, jsou nově vytvořené subjekty podrobeny mutaci (nějaké změně, například změna binárního kódu z 0 na 1). Ale tyto změny narušují původní informaci, a jsou tedy omezeny pravděpodobností (jejich vzniku). Algoritmus se opakuje, dokud nenalezne dostatečně přesnou odpověď (ukončovací podmínka) [41].

Mingyue Qiu, Yu Song a Fumio Akagi napsali v roce 2016 vědecký článek využívající kombinaci s genetickým algoritmem. Hlavním cílem bylo porovnat předpověď pomocí umělých neuronových sítí a jejich vylepšení pomocí genetického algoritmu a simulovaného žíhání. Předpovídali index Nikkei 225 z tokijské burzy z období 1993–2013. Obě kombinace vylepšily predikce umělými neuronových sítí a jako nejpřesnější byla ta s genetickým algoritmem [42]. Práci s podobnou kombinací napsali v roce 2016 i Eslam Nader Desokey, Amr Badr a Abdel Fatah Hegazy. V ní se pomocí genetického algoritmu snaží o vylepšení algoritmu shlukování metodou nejbližších středů. Avšak v této práci nejsou vstupními daty informace z burzy ale pouze zprávy z médií a sociálních sítí. Na ty byl nejdříve postaven model k-průměrů (v tomto případě 3 – nakoupit, prodat, držet) a poté genetický algoritmus který vylepšil výsledné hodnoty. Tento postup se ukázal jako efektivní a přesnost byla vyšší jak 89 % [43].

Kapitola 5

Modely předpovědi finančního trhu

5.1 Nástroje

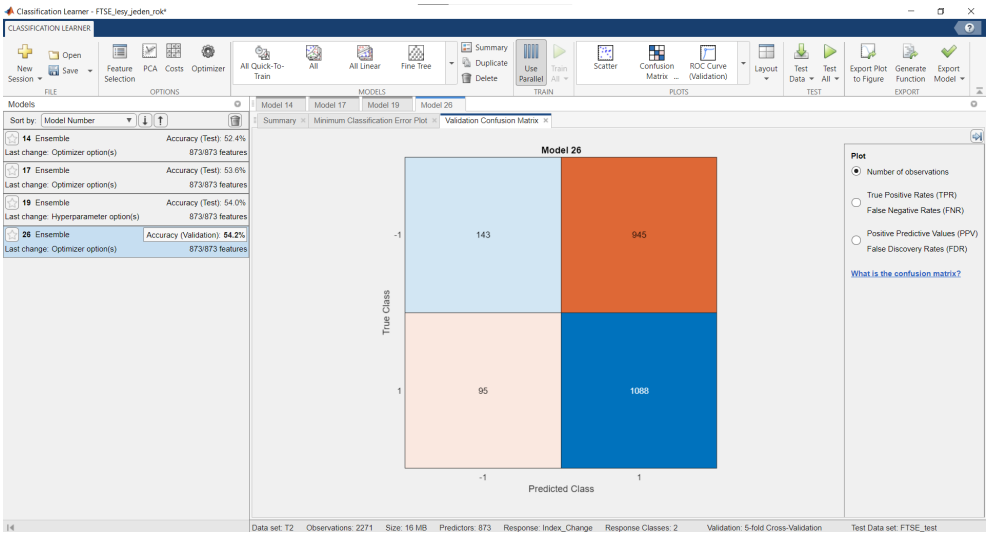
Pro strojové učení dnes existuje mnoho aplikací a knihoven. Pro modely popsané v této práci byl použít nástroje spojené programovacím jazykem Matlab a jeho prostředím. Myšlenky pro stvoření tohoto jazyku se objevují v doktorandské práci ze 60. let minulého století, kdy komerčně byl vydán až roku 1984. Software se stával populárním díky různým nástrojům vytvořených experty, která zahrnovaly velkou škálu oborů (již koncem 80. let se prodaly stovky kopií školám pro studentské účely). V rámci let se Matlab vyvíjel, například přidání knihoven pro lineární algebru v jazyku Fortran (nahrazení starších psaných v C), podpora pro grafické karty nebo také nástroje pro Strojové učení (Machine Learning and Deep learning). Ten, jak už z názvu vyplívá, obsahuje několik druhů modelů. Mezi ně patří například: neuronové sítě, hluboké učení, nebo klasifikační modely, který je používán pro modely se stromy a lesy [44].

5.1.1 Classification learner

Jak bylo zmíněno, jedná se o součást nástroje pro strojové učení. Krom stromů a lesů se zde nachází i další modely spadající do této kategorie, jako jsou například: Naivní-Bayes, metoda podpůrných vektorů nebo diskriminační analýza. Jedná se o knihovny, které se dají používat rovnou v kódu, nebo se dají ovládat pomocí dodatečného rozhraní. V něm lze nalézt uceleně všechny možnosti, které dané rozšíření poskytuje. Zásadní pro strojové učení jsou data, je zde více možností, jak je vložit pro tvoření modelů. Dají se vložit jak externě ze souborů (například v textovém formátu .csv) nebo i nahrát z interní Workspace (vnitřní paměť, kterou poskytuje prostředí Matlabu). Nemusí se tedy data nahrávat dvakrát, pokud se v předchozích krocích ještě upravují. Při tomto nahrávání se dá nastavit i křížová validace, nebo kolik dat chcete odložit na stranu pro účely pozdějšího testování. Po nahrání je programem umožněno vytváření modelů. Ty se dají vytvářen podle předem stanovených hyperparametrů, nebo se pomocí optimalizačních algoritmů hledají ty nejlepší, tedy nejlepší model.

Tyto optimalizační algoritmy jsou zde tři. Jedná se o mřížkové vyhledávání, náhodné vyhledávání a Bayesianská optimalizace. První z nich vylepšuje pomocí prohledáváním předem specifikovaných intervalů jednotlivých hyperparametrů. Vytvoří mřížku, kde se testují jednotlivé body (modely s daným podintervalem). U náhodného hledání nejsou části předefinovány a jsou vybírány náhodně všechny možné kombinace. Díky tomu může dojít k lepšímu řešení než u mřížkového, avšak je časově náročnější. Tuto vhodnou kombinaci může trvat nalézt. Za to první uvedené nalezne obecně dobré řešení, i za potřeby méně kroků. Třetí optimalizačním algoritmem je Bayeovský. Tento algoritmus generuje probabilistický model mezi hyperparametry a ověřováními cíli. Iterativně prozkoumává hyperparametry, potencionální dobré kandidáty poté vylepšuje a snaží se najít co nejvíce informací o funkci a najít optimum [45]. U těchto algoritmů se samozřejmě dají nastavit parametry (počet kroků, počet dělení, časová kvóta).

Po vytvoření modelu je zde možnost otestovat je, zde opět se mohou načíst data stejnými způsoby jako u prvotního zadávání vstupních dat (nebo je možné použít data která byla na začátku ponechána stranou). Program dále dokáže převést model na funkci, jak do jazyku Matlab, tak do C++. Nebo také ukázat různé podrobnosti modelu (např. matici záměn).



Obrázek 5.1: Rozhraní pro Classification learner.

5.2 Data

Vstupními daty jsou ceny akcií. Základní definicí je, že akcie jsou cenné papíry, které vám zaručují vlastnění části dané společnosti. S tím jsou spojená i práva na podílení se na zisku, nebo hlasování na valné hromadě akcionářů [46]. Jeho důležitou vlastností je jeho cena. Ta, a mnoho dalších, hraje roli v rozhodování burzovních makléřů – obchodníků na burze. Ty na jednotlivých trzích využívají množství typů obchodních strategií. Ty se obecně dělí na krátkodobé (short-term) a dlouhodobé (long-term).

U krátkodobých se obchoduje v rámci sekund maximálně dní. Nejkratší z nich je skalpování, kde akcie se drží pouze malé časové okno (sekundy nebo minuty) a profituje z malých výkyvů, objevujících se přes pracovní den (nejčastěji v nejvíce rušnou část dne). Většinou se obchoduje se silnějšími měnami jako je euro, dolar, libra nebo japonský jen (a pohyby mezi nimi). Dalším je například denní obchodování, kde se na trh vstupuje i vystupuje ve stejný den, tedy neponechávají se přes noc. Akcie se prodá na konci dne ať už se ziskem nebo ztrátou. Na pomezí krátkodobých a dlouhodobých je swingové obchodování. Zde se časové okno zvedá na více dnů, maximálně několik týdnů. Nejedná se o tak časově náročný způsob, je tedy hodně oblíbený mezi neprofesionálními makléři (mají obchodování jako další práci či jako koníček). Stále se však musí sledovat trh. U tohoto druhu se využívají různé taktiky jako trendové obchodování, proti trendové obchodování apod. Posledním zmíněním typem je poziční obchodování. To se zařazuje mezi dlouhodobé, se hledá co největší zisk při velkých pohybech na trhu. Časový interval se může pohybovat od několika týdnů až po roky. Obchodníci při něm využívají týdenní a měsíční stavy burzy. Pomocí technických indikátorů či fundamentálních analýz pak určují správné kupní a prodejní období [47].

Dalším ukazatelem trhu jsou tzv. indexy. Jedná se o metodu sledování, jak jistá skupina subjektů stojí na burze. Obvykle se jedná o standardizovaný pohled na výkonost koše cenných papírů v určité oblasti [48].

Ty jsou dána jako vstupní data pro modely stromů a lesů. Přesněji se jedná o indexy FTSE a S&P 500, tedy anglickou a americkou burzu. S&P index neboli Standard and Poor's 500, sleduje 500 předních firem pohybujících se na trhu ve Spojených státech. Holdingové společnosti nemovitostní akcie jsou z indexu vyřazeny. Jeho datum vzniku se připisuje k roku 1860 kde byl založena investiční informační služba Henryho Vanrum Poora. Ten se v průběhu let rozrůstal a dnešní podobu dostal v roce 1957. Na rozdíl od Dow Jones indexu se počítá vážený průměr jednotlivých akcií, a tedy silné firmy mají větší vliv na pohyb indexu. Mnoha investory je brán jako jeden z nejlepších ukazatelů výkonu trhu [49].

Druhý používaný index je obdobný S&P, kdy společnost FTSE Russell Group index tvoří jako referenční hodnoty pro světové finanční trhy. Obsahuje 100 předních firem ve Velké Británii. Vznikl v roce 1984 a od té doby jeho vlastní hodnota vzrostla na sedmináso bek. Index se pravidelně mění každých čtvrt roku (upravuje se seznam firem). Samozřejmě se upravuje i když dojde k změně mimo dané datumy uprav (například odkup společnosti a kupující je z jiného státu, slučování firem apod.) [50].

5.2.1 Formát vstupních dat

Jak již bylo zmíněno výše data se opírají o indexy S&P a FTSE. V obou případech byly použity data v období 2010–2019. Data obsahují pohyby jednotlivých firem tak i indexů. Je to děleno na Open, High, Close, Low a Volume (v případě FTSE i Adj. Close). První čtyři (a Adj.Close), popisují cenu v různých obdobích. Open je cena při otevření burzy, High její nejvyšší cena, Low její nejnižší a Close – cena

při zavření burzy (Adj. Close je v poměru k dividendám). Volume pak říká kolik se v daný den prodalo akcií. Řádek tedy koresponduje k danému dnu v roce, slupce jsou jednotlivé atributy akcií (pohyby cen a prodaných kusů), ty samé pro indexy a poslední obsahuje výstupní označení (1 pokud následující den je růst a -1 pro pokles).

U indexu S&P byly použit volně dostupný data set ze stránky Kaggle a data indexu z Google finance, který se každý den upravuje. Nebyla zde nutná tak velká úprava. Pro data z anglické burzy byly použity data z Yahoo finance, Investing.com a Google Finance. Z těchto stránek se data stahovali na základě seznamu změn indexu vytvořeným společností FTSE Russell Group a stránky stockchallenge.co.uk kde, se nacházeli historické seznamy a změny v indexu. Pomocí skriptů se potom data spojovala do jednotného souboru. V případě S&P se data pouze transformovali a byl přidány podrobnější data o pohybu indexu. Přitom to procesu se také doplňovali neúplná data (proto je využito více zdrojů) a zároveň u firem které se nepromítali do indexu dále (například firma zanikla) byly odstraněny.

Po vytvoření data setů bylo potřeba vytvořit třídu označení neboli výstup toho jednotlivého dne. Každý den byla vypočtena změna koncové ceny, tedy odečtení budoucí ceny od té stávající. Podle toho, jakého znaménka tato změna je, dojde k označení daného dne 1 nebo -1, tedy třída označení má pouze dva prvky. To znamená že pokud další den dojde k růstu, výstupním označením je 1 a naopak pro klesání.

5.2.2 Popisná statistika

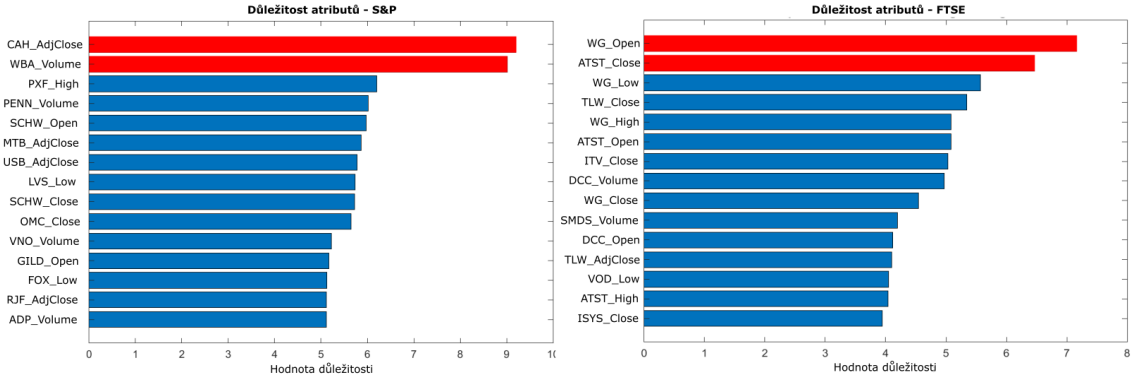
Data pro oba indexy obsahují velké množství firem (pro index FTSE jich je více jako 120 a pro S&P lehce přes 500), jsou firmy rozděleny na hlavní sektory. Jedná se o komunikační technologie, nepotřebné zboží, potřebné zboží, energie, finance, zdravotní péče, průmyslové odvětví, informační technologie, materiály, realitní kanceláře a služby. Většina kategorií je popsána již názvem (Komunikační technologie – poskytovatelé mobilních sítí atp.). Pro upřesnění tedy pouze: nepotřebné zboží jsou firmy vyrábějící produkty nebo služby, které nejsou potřeba k životu, potřebné zboží zahrnují opak, materiály jsou těžební a některý zpracovávatelský průmysl anebo že služby zahrnují poskytovatele vody, odpadních síti či plynu. Jednotlivé tabulky jsou k nalezení v příloze. V rámci indexu FTSE je nejsilnější sektor informačních technologií (z pohledu výnosnosti). Zároveň však měli jednu z největších směrodatných odchylek, tedy větší rizikovost v rámci investice. Z pohledu těchto dat tlačí dolů index FTSE sektor energií, který v průměru má zápornou výnosnost, tedy ztrátu, ale také má vysoký rozptyl. V rámci indexu S&P je situace obdobná a výnosnost má největší obor informačních technologií. To obecně opisuje obecný vývoj, kdy se jedná o rychle rostoucí sektor. Jde u něj zpozorovat i větší volatilitu. Nejhorším je opět sektor energií, avšak nemá v průměru ztrátu, ale je zde také velký rozptylem a rozdíl mezi maximem a minimem. Podrobnější informace jsou k nalezení v příloze, jak jsem zmíněno výše [51][52].

5.3 Modely a jejich výsledky

Pro oba data sety je tvořeno několik modelů. Nejprve modely stromů s pevně danými hodnotami (přednastavené), model vytvořený optimalizačním algoritmem pro stromy, AdaBoost metoda, Bag metoda a optimalizační model pro lesy. Metody AdaBoost a metoda Bag jsou vybrány jako základní reprezentativní pro náhodný les. Metoda Bag vychází z agregace bootstrapu, která je popsána výše, další metody jsou odkázány na způsob vylepšování (boosting), který je také popsán výše. Pro všechny modely byla využita pětinásobná křížová validace.

5.3.1 Důležitost atributů

Jednotlivým atributům se dá přiřadit určitá významnost, kterou určí algoritmus. Tedy jak důležitý je daný atribut pro charakteristiku data setu. Pro tento účel byl vybrán algoritmus Chi-kvadrát, a to pro oba indexy. U indexu S&P algoritmus rozpoznal jako atributy z velkou mírou důležitosti 2 atributy, respektive s nejvyššími hodnotami skóre hodnocení. První byl CAH_AdjClose tedy upravená koncová cena akcie s označením CAH, hodnota byla 9,2036. Druhý atribut s podobným výsledkem byl WBA_Volume, to znamená množství prodaných akcií WBA. Další atributy začínají na hodnotách pohybujících se okolo 6 a pomalu klesají. U FTSE indexu není situace tak jednoznačná neboli že není tak velký rozdíl mezi prvními několika místy a dalšími místy v prvních 15. Prvním je zde Wg_Open (cena akcie WG při otevření burzy) s hodnotou 7,1632, druhý je ATST_Close s 6,4658 a třetí je WG_Low s 5,5720. Se snižováním příček se snižuje i skóre důležitosti, avšak se zpomalujícím tempem. Dalším poznatkem je velké zastoupení akcií WG a ATST v prvních 15 příčkách.



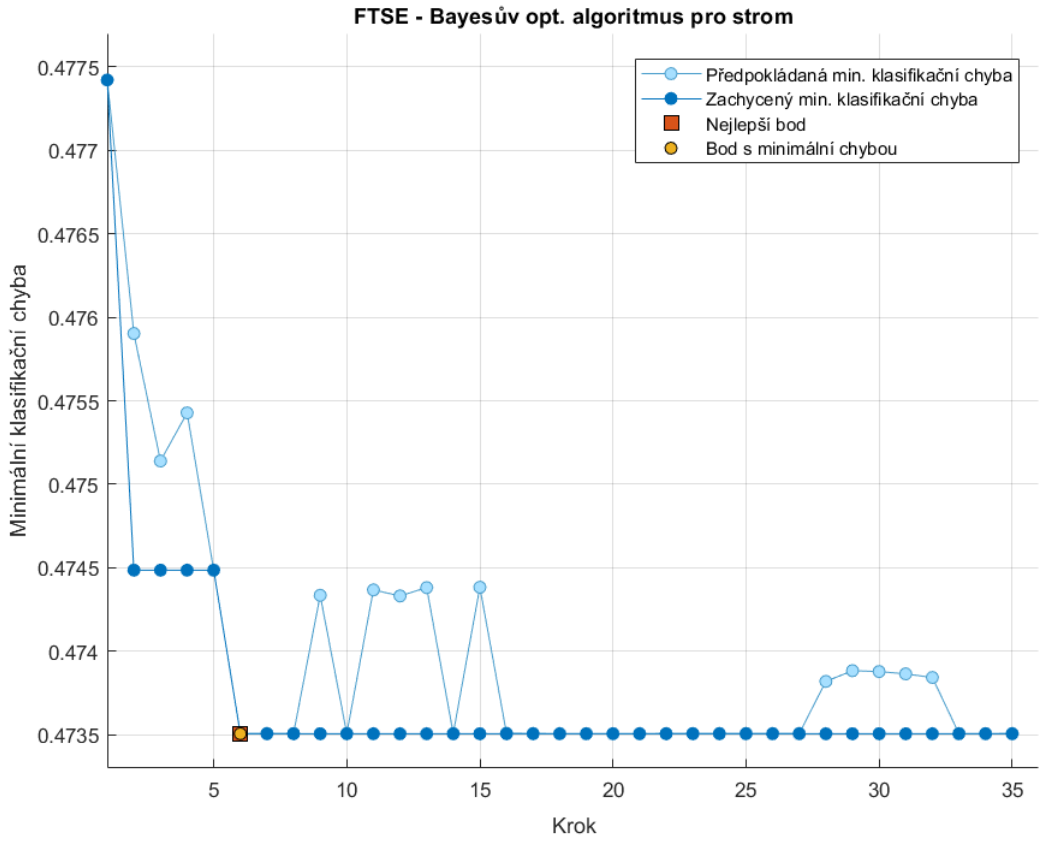
Obrázek 5.2: Grafy prvních 15 atributů pro indexy S&P a FTSE, červeně jsou označeny vždy první dva nejlepší atributy

5.3.2 Modely s daty FTSE indexu

Jak bylo zmíněno výše, data jsou z období 2010-2019. Prvních osm let se používají jako tréninková data a poslední rok jako testovací. Nejprve jsme začali tvorbou

stromů, Ty jsou rozděleny na tři před vytvořené šablony a poté na vznik stromového modelu pomocí optimalizačního algoritmu. Celkový počet atributů je 897, kde atributy jsou jednotlivé ceny akcií nebo počet prodaných.

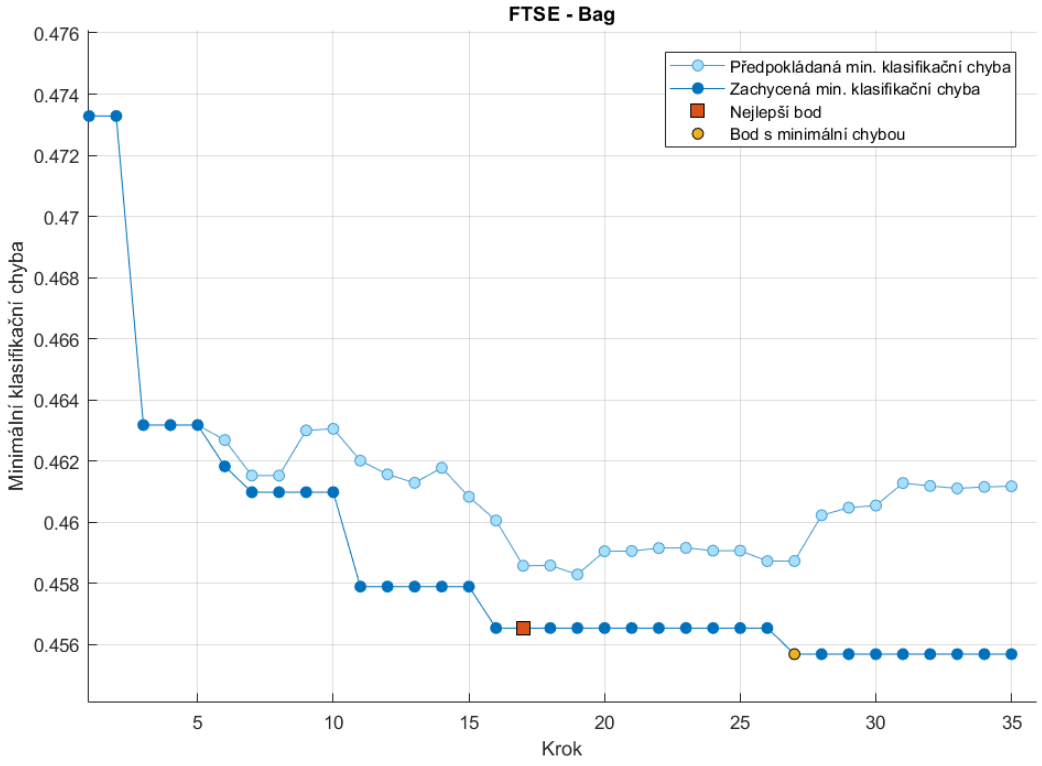
Tři předdefinované jsou Coarse tree, Medium tree a Fine tree. Ty všechny používají Gini index jako rozhodující pravidlo větvení, odlišují se jeho počtem. První z nich má za hyperparametr větvení číslo 4, Medium tree má 20 a poslední je nejsložitější s počtem 100. U těchto stromů byla vidět postupná klesající tendence v přesnosti s rostoucím počtem maximálního větvení kdy u nejjednoduššího stromu dosahovala přesnost 52,3 % po validaci a 52 % při testování, u středního stromu byly výsledky 51,6 % a 51 %, u třetího stromu 51,7 % a 50,8 %.



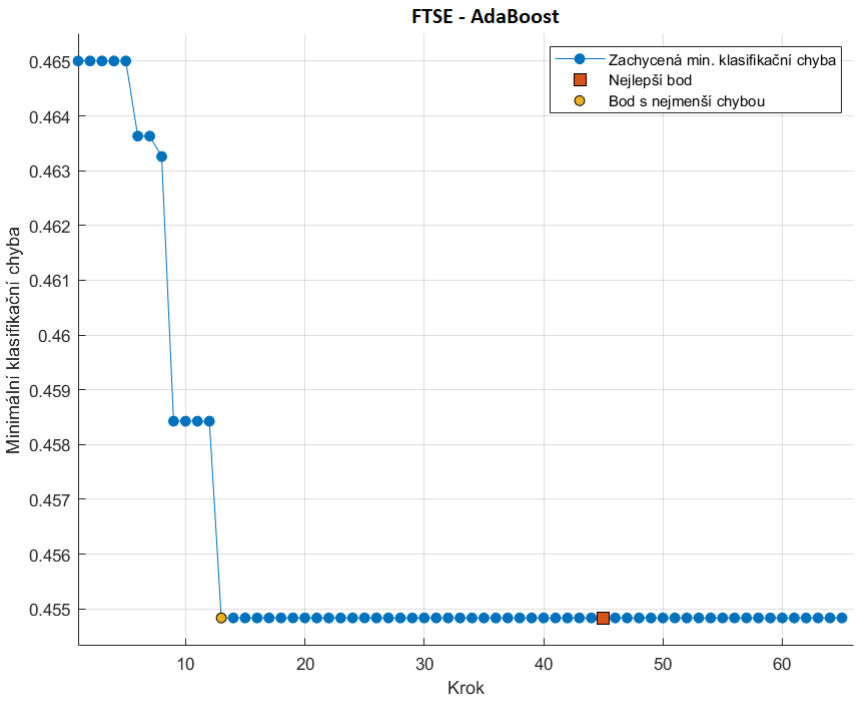
Obrázek 5.3: Graf vývoje stromu pomocí Bayesovi optimalizace

Další se tvořil strom pomocí optimalizační algoritmu, který hledal nejlepší hyperparametry. Používána byla Bayesova optimalizace. To je dobré pro jednodušší struktury a dokáže v nich najít řešení poměrně rychle. Pro parametr bylo zvoleno 30, avšak algoritmus nalezl nejlepší hyperparametry už při 6. kroku (naproti tomu jeden z horších případu náhodného hledání nalezl až po 20. kroku hledání nejlepší řešení, viz Obrázek 5.3). Jako nejlepší shledal algoritmus, že strom má omezení na maximálně 2 větvení a jako rozhodovací pravidlo se používá Gini index. Dosáhl vyšší přesnosti než předem zmíněné stromy, kdy přesnost validace i testování byla 52,6 %.

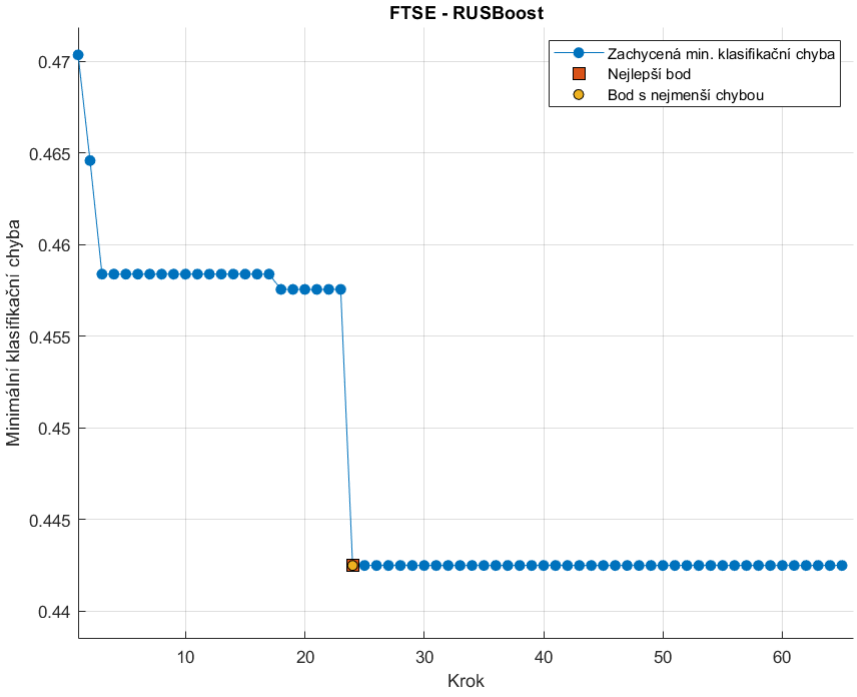
Dalším jsou souborové metody, v tomto případě lesy. Jedná se o tři modely – Bag, AdaBoost a RUSBoost. Pro všechny byli nalezeny optimální hyperparametry pomocí optimalizačního algoritmu. Pro Bag metodu byla využita Bayesova optimalizace s 35 kroky. V tomto případě se objevily 2 možnosti. První s nejlepšími vlastnostmi ohledem na hyperparametry (Nejlepší bod) a bod s nejmenší chybou. Jako nejlepší byl vybrán model s 254 studenty (stromy) a maximální počet dělení je 2154. Při validaci vyšla přesnost 53,9 % a u testování 52.4 %. Druhý model měl nejlepší výsledky. U něj bylo pro porovnání použito náhodné hledání (Bayesova se dopracovala stejného výsledku, pouze o několik kroků dříve). Jako výsledné hyperparametry vybrala optimalizace, že počet stromů je 77, míra učení je 0,0479 a maximální počet dělení je 187 (u nejlepšího bodu). Přesností překonává oba modely, měl 54,2 % u validace a 54 % u testování. Poslední model byl vybrán náhodným hledáním. Tomu vyšla nejlépe metoda RUSBoost, s parametry: 158 počet stromů, 0,17 míra učení a 51 maximálních větvení. Přesnost u validace vyšla podobně jako u předchozího modelu a to 54 % ale u testovaní klesla na 53,6 %.



Obrázek 5.4: Graf vývoje metody Bag u FTSE indexu



Obrázek 5.5: Graf vývoje metody Adaboost pro index FTSE



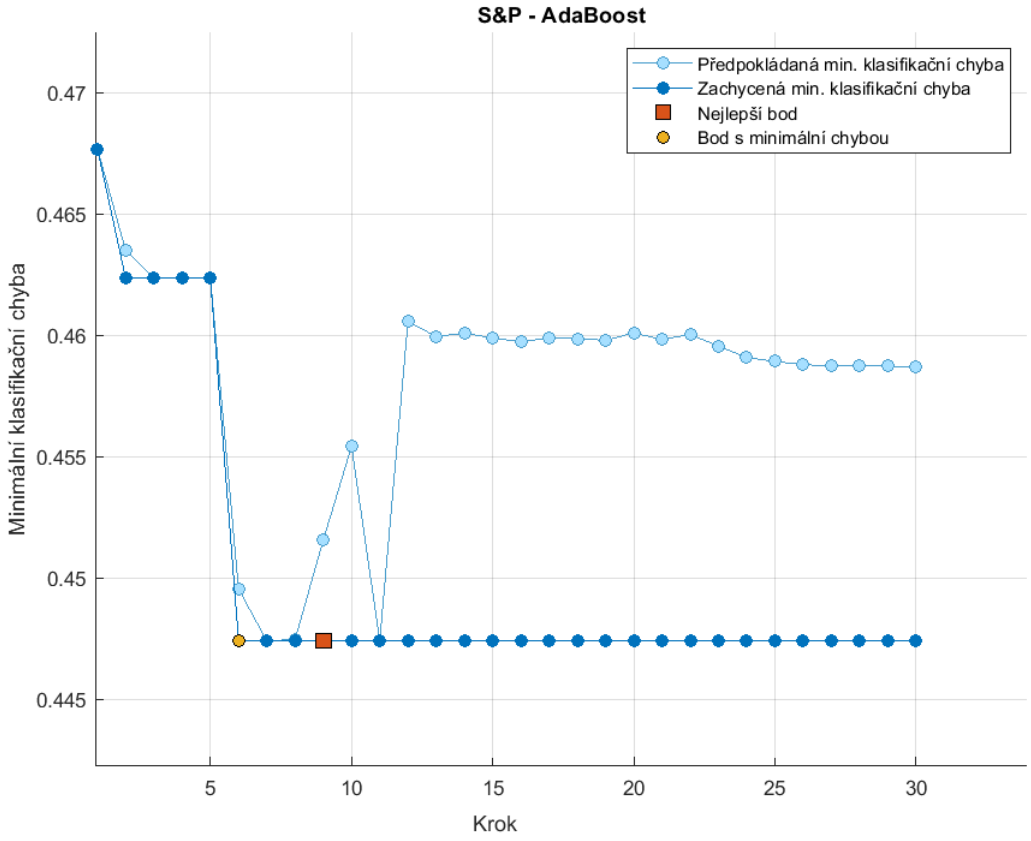
Obrázek 5.6: Graf vývoje optimalizace při metodě RUSBoost

5.3.3 Model s daty S&P indexu

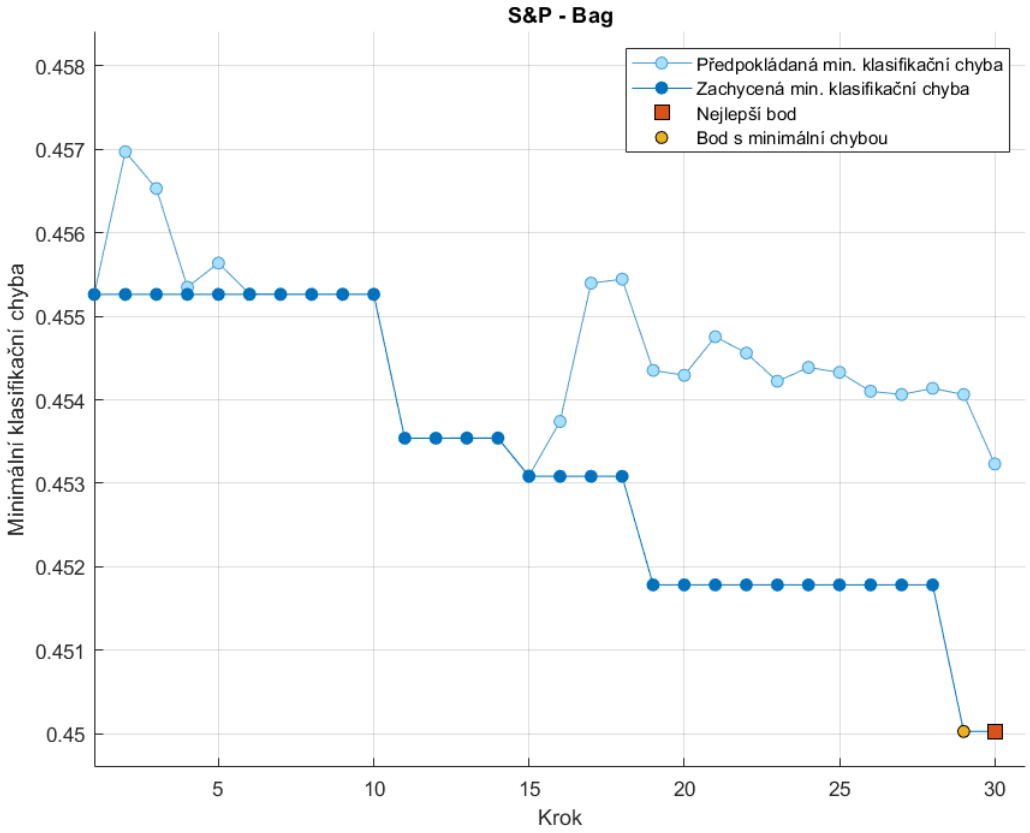
U těchto modelů je nastavení stejné. Dělení osm let ku jednomu roku, tři stromy ale pouze 2 souborové metody, jelikož optimalizace vybrala jako nejlepší metodu Bag. Data set obsahuje násobně více atributů, jelikož se zde objevuje i více subjektů. Objevuje se jich zde 2987.

Opět byly použity tři základní předdefinované modely pros stromy. První model s maximálním větvením 2 mělo přesnost validace 54,1 % a testování 43,9 %. Model Medium tree se dopracoval k výsledku 51,9 % při validaci a 42,2 %. Třetí a nejsložitější z uvedených stromů měl přesnost 52,2 % u validace a 43,9 % u testování. Obdobně na tom byla i optimalizace, u které vyšly přesnosti 54,2 % a 41,8 %. Tedy žádný z těchto jednodušších modelů nepřekonal 50 % hranici pro testovací soubor.

Dalšími jsou tedy dva modely lesů. Výsledky byli lepší než u předchozích stromů. Nejdříve metoda Adaboost, kterou vyprodukovala i optimalizace. Přesnost se zde ustálila na 54,7 % u validace a 51,8 % u testování. Jeho hyperparametry byly maximálně 2 větvení, míra učení 0,01 a počet stromů byl 500. U metody Bag byla přesnost (u testování) vyšší. Přesněji 54,4 % a 52 %. S toho lze vypozorovat, že pro tento větší data set lépe odhadly tyto složitější modely.



Obrázek 5.7: Graf vývoje optimalizace metody Adaboost a indexu S&P



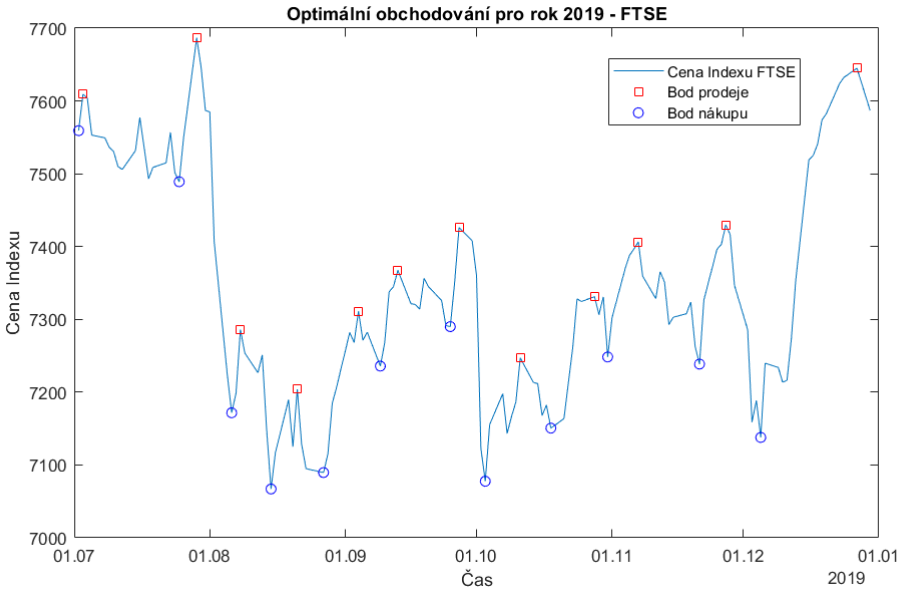
Obrázek 5.8: Graf vývoje optimalizace pro metodu Bag a index S&P

5.4 Interpretace výsledků

Obecné přesnosti byli už zmíněny v kapitole výše, ukážeme tedy jak lze tyto modely přetvořit na praktické využití. Použijeme tedy 2.polovinu roku 2019 jako obchodovací období. Zjistíme předpověď a porovnáme s reálnými hodnotami.

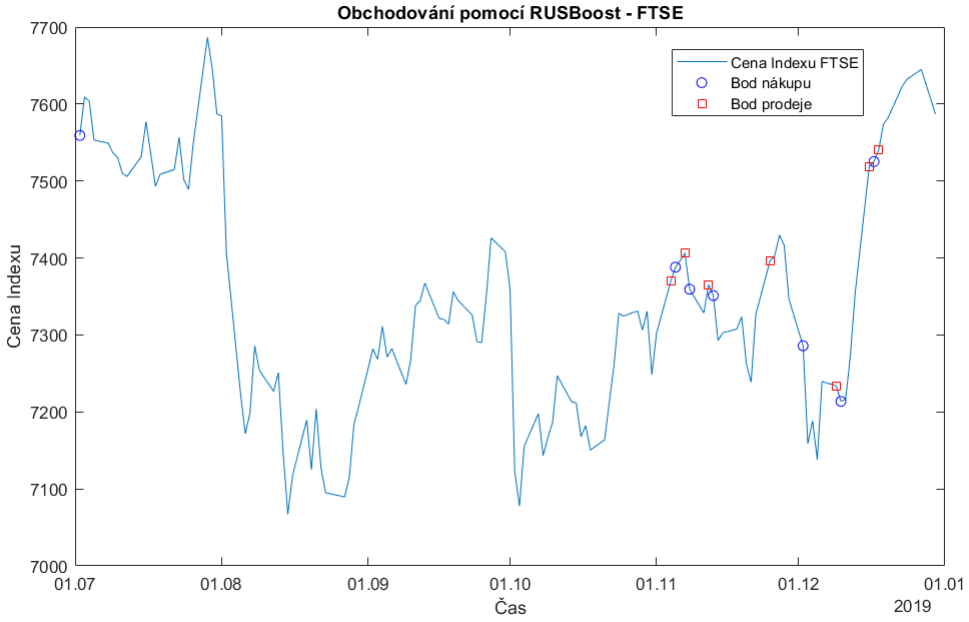
Obchodování s FTSE indexem

Nejprve budeme předpovídat index FTSE. I když model s použitím AdaBoostu, tak v matici záměn má však vysoké TPR (True Positive Rates) a ve výsledném půlroku nedokáže odhadnou poklesy v ceně. Použijeme tedy RUSBoost a Bag a porovnáme který bude mít lepší výsledky. Porovnáváme tedy optimální obchodování (tedy pokud zareaguje na změny, a ne dojde ke ztrátě a k co největšímu zisku, i za předpokladu že víme co se v budoucnu stane), za pomocí metody Bag a RUSBoost.

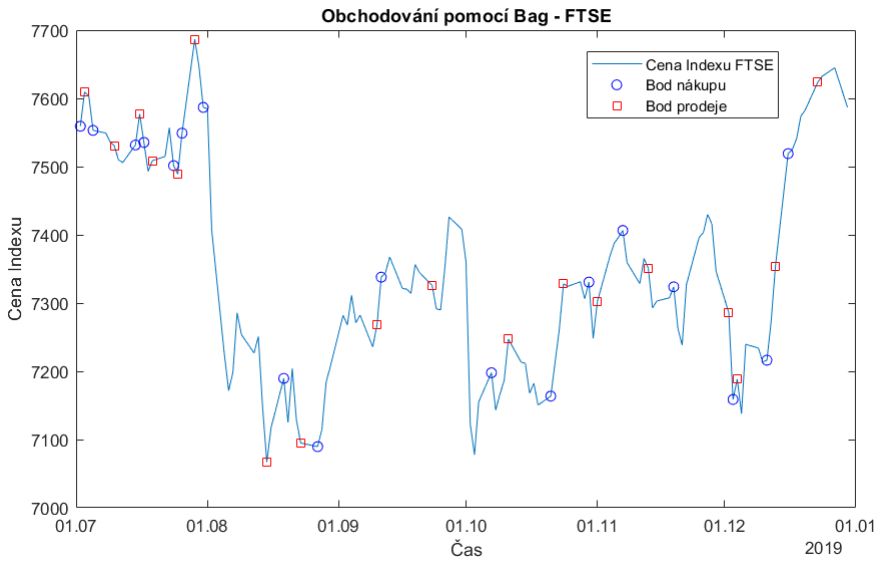


Obrázek 5.9: Graf s vyznačenými body optimálního obchodování u indexu FTSE

Na Obrázku 5.9 můžeme vidět graf vývoje ceny indexu FTSE v 2. polovině roku 2019 s vyznačenými body nákupu a prodeje. Body se vybírali za určité míry přesnosti, u jak velkého kolísaní má dojít k nákupu, příp. prodeji. Tyto body budou porovnávány s body, které vybraly metody RUSBoost a Bag. Ty byly vybrány z výstupních dat, které ukazovali pouze zdali je předpokládán růst, nebo pád. Proto byly vždy vybrány body korespondující s předpovídanou akcí. Tedy nákup byl vybrán, pokud měla cena stoupat a prodej, pokud měla klesat.



Obrázek 5.10: Graf vývoje ceny indexu FTSE s body generovanými z výstupu modelu RUSBoost



Obrázek 5.11: Graf vývoje ceny indexu FTSE s body generovanými z výstupu modelu Bag

Z těchto 2 modelů vyprodukoval více bodů metoda Bag a to 18 dvojic bodů, kde RUSBoost měl pouze 7. Optimální obchodování má 12 dvojic. Nejlepší výsledek má samozřejmě optimální, a to 2194,87. Metoda RUSBoost vyprodukovala většinu bodu při konci roku ale i tak dokázal dojít kladného výsledku 148,69. Poslední v poměru k výdělkům byla metoda Bag. Ta měla více bodů, ale odhadla největší pokles špatně a tím si pohoršila ve výsledném součtu o -519,77. Stále ale zůstala v kladné rovině a utržila 87,38.

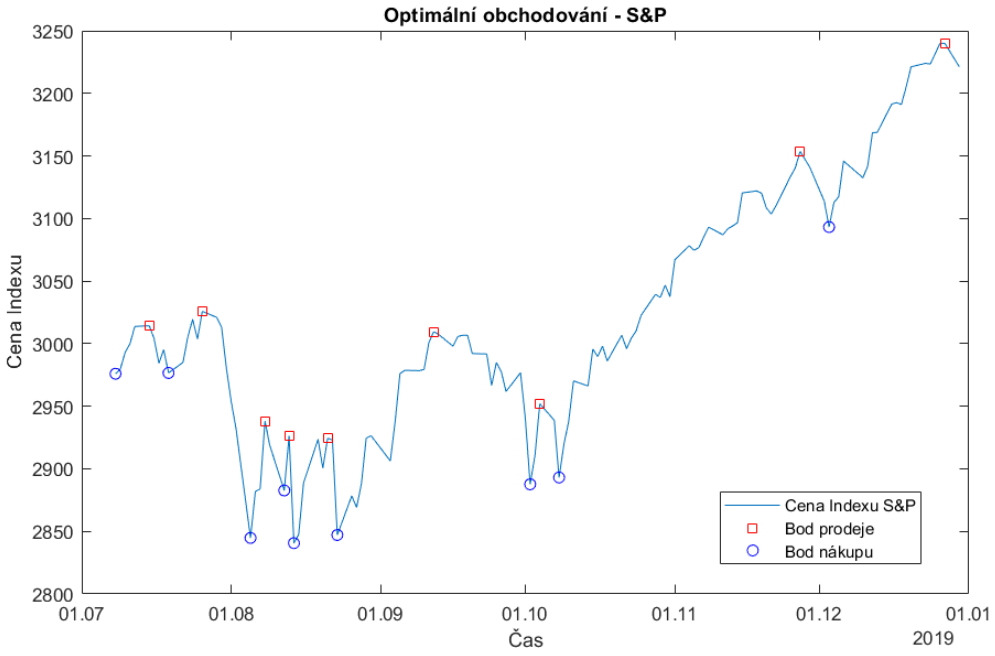
Optimální obchodování			Metoda Bag			Metoda RUSBoost		
Nákupní cena	Prodejní cena	Rozdíl	Nákupní cena	Prodejní cena	Rozdíl	Nákupní cena	Prodejní cena	Rozdíl
7559,19	7609,32	50,13	7559,19	7609,32	50,13	7559,19	7369,69	-189,50
7489,05	7686,61	197,56	7553,14	7530,69	-22,45	7388,08	7406,41	18,33
7171,69	7285,90	114,21	7531,72	7577,20	45,48	7359,38	7365,44	6,06
7067,01	7203,97	136,96	7535,46	7508,70	-26,76	7351,21	7396,29	45,08
7089,58	7311,26	221,68	7501,46	7489,05	-12,41	7285,94	7233,90	-52,04
7235,81	7367,46	131,65	7549,06	7686,61	137,55	7213,76	7519,05	305,29
7289,99	7426,21	136,22	7586,78	7067,01	-519,77	7525,28	7540,75	15,47
7077,64	7247,08	169,44	7189,65	7094,98	-94,67			
7150,57	7331,28	180,71	7089,58	7267,95	178,37			
7248,38	7406,41	158,03	7338,03	7326,08	-11,95			
7238,55	7429,78	191,23	7197,88	7247,08	49,20			
7137,85	7644,90	507,05	7163,64	7328,25	164,61			
			7330,78	7302,42	-28,36			
			7406,41	7351,21	-55,20			
			7323,80	7285,94	-37,86			
			7158,76	7188,50	29,74			
			7216,25	7353,44	137,19			
			7519,05	7623,59	104,54			
Součet		2194,87	Součet		87,38	Součet		148,69

Tabulka 5.1: Tabulka výsledků obchodování pro index FTSE

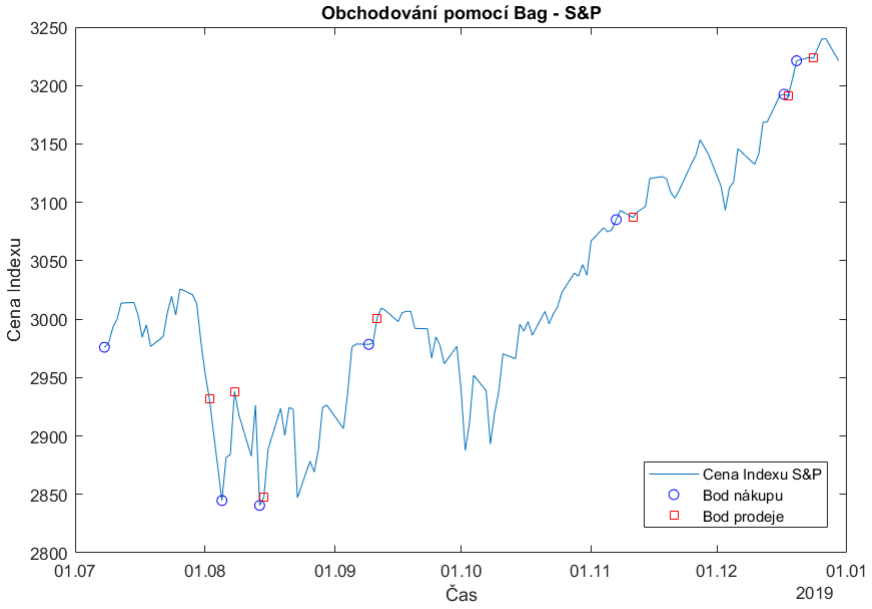
5.4.1 Obchodování s S&P indexem

Dále tedy máme modely pro indexy S&P. Zde použijeme také dva, a to Adaboost a Bag, protože předčily v přesnosti ostatní modely o desítky procent a zároveň není zde problém s přehnaným určováním jednoho atributu výstupní třídy. Opět porovnáme s optimálním řešením.

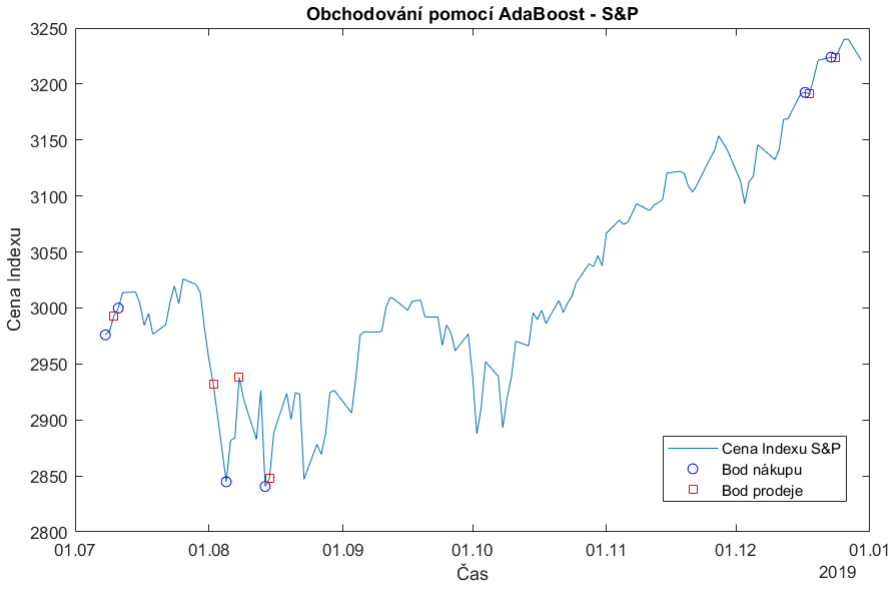
Graf na Obrázku 5.12 ukazuje, jak postupovala cena ve sledovaném období i s vyznačenými body. Oproti FTSE není tolik kolísavý a neobsahují žádný velký pád, proto i bodů nákupu a prodejů není tolik.



Obrázek 5.12: Graf vývoje ceny indexu S&P a optimálních bodů pro obchodování



Obrázek 5.13: Graf vývoje ceny indexu S&P a body vygenerované metodou Bag



Obrázek 5.14: Graf vývoje ceny indexu S&P a body generované metodou AdaBoost

Pro oba modely byly vybrány body z výstupu jejich modelů podle stejných pravidel jako u FTSE indexu. Nejméně bodů má Adaboost, 6 dvojic. Zároveň jsou více rozpoloženy po okrajích obdobích. Les vytvoření pomocí metodou Bag má 7 dvojic bodů. Referenční optimální strategie měla celkový součet 942,65, a násobně převyšoval výsledky predikcí. Avšak obě se udrželi v kladných číslech. Adaboost skončil s 81,56 a Bag s 47,6. Tedy byly úspěšné ve sledovaném období, ale stále nedokázali rozpoznat některé větší změny, anebo velký propad zaměnily za nárůst a prodali

se ztrátou. Avšak díky odhadů některých menších změn dokázali zůstat v kladné hladině.

Optimální obchodování			Metoda Bag			Metoda AdaBoost		
Nákupní cena	Prodejní cena	Rozdíl	Nákupní cena	Prodejní cena	Rozdíl	Nákupní cena	Prodejní cena	Rozdíl
2975,95	3014,3	38,35	2975,95	2932,05	-43,9	2975,95	2993,07	17,12
2976,61	3025,86	49,25	2844,74	2938,09	93,35	2999,91	2932,05	-67,86
2844,74	2938,09	93,35	2840,6	2847,6	7	2844,74	2938,09	93,35
2882,7	2926,32	43,62	2978,43	3000,93	22,5	2840,6	2847,6	7
2840,6	2924,43	83,83	3085,18	3087,01	1,83	3192,52	3191,14	-1,38
2847,11	3009,57	162,46	3192,52	3191,14	-1,38	3224,01	3223,38	-0,63
2887,61	2952,01	64,4	3221,22	3223,38	2,16			
2893,06	3153,63	260,57						
3093,2	3240,02	146,82						
Součet		942,65	Součet		81,56	Součet		47,6

Tabulka 5.2: Tabulka výsledků obchodování pro index S&P

Závěr

Cílem této práce bylo vytvoření modelů predikujících finanční trh, kdy bylo vybráno předpovídání vývoje indexů FTSE a S&P.

Nejprve jsem se seznámil s teorií, která leží na pozadí strojového učení. Speciálně se práce zaměřuje na stromové metody a souborové metody k nim přidružené. Dále jsem se seznámil z dalšími pracemi na téma predikce trhu pomocí strojového učení. Následovalo seznámení se s použitými nástroji programu Matlab. Po těchto krocích následovalo tvoření modelů.

Prvním důležitým krokem pro tvoření strojového učení je výběr dat. Kdy jsem vybral indexy S&P a FTSE, jelikož se jedná o jedny z nejsilnější ukazatelů trhu. První v Spojených státech a druhý ve Velké Británii. S těmito daty jsem vytvořil několik modelů, a to jak samostatné stromy, tak i souborové metody. V rámci interpretace byly ty nejlepší byly vybrány a bylo s nimi provedeno obchodování jejich výsledky byly ukázány v porovnání s optimálními body obchodování.

Nejlepší modely pro oba indexy přesáhly 50 % hranici. Ale ani to že měl model nejlepší výsledky neznamenalo, že se jedná o vhodný model. Příkladem tohoto modelu byl využívající AdaBoost s daty FTSE indexu. I když měl přesnost 54 % tak ale jeho matice záměn nebyla vyvážená a většinu dnů odhadl na stejnou hodnotu. Dalším poznatkem je, že i když přesnost nepřesáhla 55 % tak v obchodovacím úseku nezaznamenali ztrátu. Lze ale rozpoznat, jedná-li se o velké změny, například rychlé pády, tak je model klasifikoval jako růst (příkladem je velký pokles ceny u indexu FTSE a špatné rozpoložení bodů na tomto úseku). Některé však ani nedokázali tyto změny ani zaregistrovat.

Vylepšováním těchto nedostatků by se mohli zabývat další práce. Ty by mohli přidat další makroekonomické ukazatele jednotlivým firmám, jako se používá u různých analýz. V současnosti také roste vliv sociálních sítí a medií na ceny akcií, takže by se k datům z trhů mohli přidat i sentimentální analýza. Ta by mohla odhalit některé velké změny (příkladem mohou být různé kauzy, zvěsti nebo i oficiální oznámení firmy).

Literatura

[1] HAENLEIN, Michael a Andreas KAPLAN. *A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. California Management Review*. 2019, 61(4), 5-14. ISSN 0008-1256. Dostupné z: doi:10.1177/0008125619864925

[2] *Artificial intelligence, n.*. Oxford English Dictionary [online]. Oxford University Press [cit. 2022-01-13]. Dostupné z: `https://www.oed.com/viewdictionaryentry/Entry/271625`

[3] KOK, Joost N., et al. *Artificial intelligence: definition, trends, techniques, and cases*. Artificial intelligence, 2009, 1: 270-299.

[4] *Turing test*. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, [cit. 2022-01-13]. Dostupné z: `https://en.wikipedia.org/wiki/Turing_test`

[5] H. WITTEN, Ian, Eibe FRANK, Mark A. HALL a Christopher PAL. *Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann Series in Data Management Systems)*. 4th Edition. Morgan Kaufmann, [2016]. ISBN 978-0128042915.

[6] *Machine learning*. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2022-01-13]. Dostupné z: `https://en.wikipedia.org/wiki/Machine_learning`

[7] MURPHY, Kevin P. *Machine learning: a probabilistic perspective*. Cambridge: MIT Press, 2012. Adaptive computation and machine learning series. ISBN 978-0-262-01802-9.

[8] *Regression Analysis in Machine learning*. JavaTpoint [online]. Noida [cit. 2022-01-22]. Dostupné z: `https://www.javatpoint.com/regression-analysis-in-machine-learning`

[9] *Regression analysis*. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation [cit. 2022-01-22]. Dostupné z: `https://en.wikipedia.org/wiki/Regression_analysis`

[10] MISHRA, Sanatan. *Unsupervised Learning and Data Clustering*. Towards Data Science [online]. Medium, 2017 [cit. 2022-01-22]. Dostupné z: `https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a`

[11] *Semi-supervised learning.* In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation [cit. 2022-01-22]. Dostupné z: https://en.wikipedia.org/wiki/Semi-supervised_learning

[12] *Information gain ratio.* In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation [cit. 2022-01-22]. Dostupné z: https://en.wikipedia.org/wiki/Information_gain_ratio

[13] *Reinforcement learning.* In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation [cit. 2022-01-30]. Dostupné z: https://en.wikipedia.org/wiki/Reinforcement_learning

[14] ALPAYDIN, Ethem. *Introduction to Machine Learning.* Third Edition. Massachusetts: The MIT Press, 2014. ISBN 978-0-262-02818-9.

[15] *Machine Learning 4. Decision Trees.* Stiftung Universität Hildesheim [online]. Hildesheim, 2007 [cit. 2022-02-05]. Dostupné z: https://www.ismll.uni-hildesheim.de/lehre/ml-07w/skript/ml-2up-04-decisiontrees.pdf

[16] *Bootstrap aggregating.* In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation [cit. 2022-02-05]. Dostupné z: https://en.wikipedia.org/wiki/Bootstrap_aggregating

[17] *File:Ensemble Bagging.svg.* In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation [cit. 2022-02-05]. Dostupné z: https://en.wikipedia.org/wiki/File:Ensemble_Bagging.svg

[18] BREIMAN, Leo a Andreas KAPLAN. *A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence.* Machine Learning. 2019, 45(1), 5-32. ISSN 08856125. Dostupné z: doi:10.1023/A:1010933404324

[19] KOMPRDOVÁ, Klára. *Rozhodovací stromy a lesy.* Brno: Akademické nakladatelství CERM, 2012. ISBN 978-80-7204-785-7.

[20] TAHSILDAR, Shagufta. *Gini Index: Decision Tree, Formula, and Coefficient.* Quant Insti [online]. 2019 [cit. 2022-02-25]. Dostupné z: https://blog.quantinsti.com/gini-index/

[21] ROKACH, Lior. *Ensemble-based classifiers: On the Past, Present, and Future of Artificial Intelligence. Artificial Intelligence Review.* 2010, 33(1-2), 1-39. ISSN 0269-2821. Dostupné z: doi:10.1007/s10462-009-9124-7

[22] ZHOU, Zhi-Hua. *Ensamble Methods: Foundation and Algorithms.* Verze: 20120501. Roca Baton: CRC Press, 2012. ISBN 978-0-262-02818-9.

[23] WOODRUFF, Katherine. *Introduction to boosted decision trees.* Indico [online]. New Mexico State University, 2017 [cit. 2022-03-05]. Dostupné z: https://indico.fnal.gov/event/15356/contributions/31377/attachments/19671/24560/DecisionTrees.pdf

[24] *Applications of Machine learning.* JavaTpoint [online]. Nodia [cit. 2022-03-08]. Dostupné z: https://www.javatpoint.com/applications-of-machine-learning

[25] SEGAL, Troy. *Fundamental Analysis.* Investopedia [online]. Dotdash Meredith, 2021 [cit. 2022-03-08]. Dostupné z: https://www.investopedia.com/terms/f/fundamentalanalysis.asp

[26] HAYES, Adam. *Technical Analysis.* Investopedia [online]. Dotdash Meredith, 2021 [cit. 2022-03-08]. Dostupné z: https://www.investopedia.com/terms/t/technicalanalysis.asp

[27] DONGARE, A.D., R.R. KHARDE a Amit D. KACHARE. *Introduction to Artificial Neural Network. International Journal of Engineering and Innovative Technology.* 2010, 2(1), 1-6. ISSN 2277-3754. Dostupné také z: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1082.1323&rep=rep1&type=pdf

[28] GURESEN, Erkam, Gulgun KAYAKUTLU a Tugrul U. DAIM. *Using artificial neural network models in stock market index prediction. Expert Systems with Applications.* 2011, 38(8), 10389-10397. ISSN 09574174. Dostupné z: doi:10.1016/j.eswa.2011.02.068

[29] BING, Yang, Jian Kun HAO a Si Chang ZHANG. *Stock Market Prediction Using Artificial Neural Networks. Advanced Engineering Forum.* 2012, 6-7(8), 1055-1060. ISSN 2234-991X. Dostupné z: doi:10.4028/www.scientific.net/AEF.6-7.1055

[30] MOGHADDAM, Amin Hedayati, Moein Hedayati MOGHADDAM a Morteza ESFANDYARI. *Stock market index prediction using artificial neural network.* Journal of Economics, Finance and Administrative Science. 2016, 21(41), 89-93. ISSN 20771886. Dostupné z: doi:10.1016/j.jefas.2016.07.002

[31] CORTES, Corinna a Vladimir VAPNIK. *Suport-Vector Networks.* Kluwer Academic Publisher. Boston, 1995, 20, 273-297. Dostupné také z: http://image.diku.dk/imagecanon/material/cortes_vapnik95.pdf

[32] SHEN, Shunrong; JIANG, Haomiao; ZHANG, Tongda. *Stock market forecasting using machine learning algorithms.* Department of Electrical Engineering, Stanford University, Stanford, CA, 2012, 1-5.

[33] LUO, Linkai a Xi CHEN. *Integrating piecewise linear representation and weighted support vector machine for stock trading signal prediction.* Applied Soft Computing. Boston, 2013, 13(2), 806-816. ISSN 15684946. Dostupné z: doi:10.1016/j.asoc.2012.10.026

[34] PAN, Yuchen, Zhi XIAO, Xianning WANG a Daoli YANG. *A multiple support vector machine approach to stock index forecasting with mixed frequency sampling.* Knowledge-Based Systems. Boston, 2017, 122(2), 90-102. ISSN 09507051. Dostupné z: doi:10.1016/j.knosys.2017.01.033

[35] CHONG, Eunsuk, Chulwoo HAN, Frank C. PARK a Daoli YANG. *Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies.* Expert Systems with Applications. Boston, 2017, 83(2), 187-205. ISSN 09574174. Dostupné z: doi:10.1016/j.eswa.2017.04.030

[36] ZHONG, Xiao a David ENKE. *Predicting the daily return direction of the stock market using hybrid machine learning algorithms.* Financial Innovation. Boston, 2019, 5(1), 187-205. ISSN 2199-4730. Dostupné z: doi:10.1186/s40854-019-0138-0

[37] PATEL, Jigar, Sahil SHAH, Priyank THAKKAR a K KOTECHA. *Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques.* Expert Systems with Applications. Boston, 2015, 42(1), 259-268. ISSN 09574174. Dostupné z: doi:10.1016/j.eswa.2014.07.040

[38] UMER, Muhammad, Muhammad AWAIS a Muhammad MUZAMMUL. *Stock Market Prediction Using Machine Learning(ML)Algorithms.* ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal. Boston, 2020, 8(4), 97-116. ISSN 2255-2863. Dostupné z: doi:10.14201/ADCAIJ20198497116

[39] Javed Awan, Mazhar and Mohd Rahim, Mohd Shafry and Nobanee, Haitham and Nobanee, Haitham and Munawar, Ashna and Yasin, Awais and Zain, Azlan Mohd. *Social Media and Stock Market Prediction: A Big Data Approach.* M. J. Awan, M. Shafry, H. Nobanee, A. Munawar, A. Yasin et al., "Social media and stock market prediction: a big data approach,"Computers, Materials & Continua, vol. 67, no.2, pp. 2569–2583, 2021, Dostupné na SSRN: `https://ssrn.com/abstract=3827106`

[40] KHAN, Wasiat, Mustansar Ali GHAZANFAR, Muhammad Awais AZAM, Amin KARAMI, Khaled H. ALYOUBI a Ahmed S. ALFAKEEH. *Stock market prediction using machine learning classifiers and social media, news.* Journal of Ambient Intelligence and Humanized Computing. Boston, 2022, 13(7), 3433-3456. ISSN 1868-5137. Dostupné z: doi:10.1007/s12652-020-01839-w

[41] LUNER, Petr. Jemný úvod do genetických algoritmů. *Computer Graphics Charles University* [online]. Praha [cit. 2022-03-27]. Dostupné z: `https://cgg.mff.cuni.cz/~pepca/prg022/luner.html`

[42] QIU, Mingyue, Yu SONG a Fumio AKAGI. *Application of artificial neural network for the prediction of stock market returns: The case of the Japanese stock market.* Journal of Ambient Intelligence and Humanized Computing. Boston, 2016, 85(7), 1-7. ISSN 09600779. Dostupné z: doi:10.1016/j.chaos.2016.01.004

[43] DESOKEY, Eslam Nader, Amr BADR a Abdel Fatah HEGAZY. *Enhancing stock prediction clustering using K-means with genetic algorithm.* 2017 13th International Computer Engineering Conference (ICENCO). Boston: IEEE, 2017, 2017, 85(7), 256-261. ISBN 978-1-5386-4266-5. ISSN 09600779. Dostupné z: doi:10.1109/ICENCO.2017.8289797

[44] *MATLAB*. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation [cit. 2022-06-20]. Dostupné z: https://en.wikipedia.org/wiki/MATLAB

[45] BERGSTRA, James, et al. *Algorithms for hyper-parameter optimization. Advances in neural information processing systems*, 2011, 24.

[46] *Akcie*. UniCredit Bank [online]. Praha, 2017 [cit. 2022-06-28]. Dostupné z: https://www.unicreditbank.cz/content/dam/cee2020-pws-cz/cz-dokumenty-2017/obcane/mojecile/Akcie_CZ.pdf

[47] *Different Types Of Strategies*. Capital Index [online]. Londýn [cit. 2022-06-28]. Dostupné z: https://www.capitalindex.com/uk/eng/pages/trading-guides/different-types-of-trading-strategies

[48] CHEN, James. *Index*. Investopedia [online]. Dotdash Meredith, 2021 [cit. 2022-06-28]. Dostupné z: https://www.investopedia.com/terms/i/index.asp

[49] *S&P 500*. Britannica [online]. Encyclopædia Britannica [cit. 2022-06-28]. Dostupné z: https://www.britannica.com/topic/SandP-500

[50] YOUNG, Julie. *Financial Times Stock Exchange Group (FTSE)*. Investopedia [online]. Dotdash Meredith, 2021 [cit. 2022-06-28]. Dostupné z: https://www.investopedia.com/terms/f/ftse.asp

[51] *List of S&P 500 companies*. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation [cit. 2022-07-01]. Dostupné z: https://en.wikipedia.org/wiki/List_of_S%26P_500_companies

[52] *FTSE 100 Index Sector Weightings*. Siblis Research [online]. Helsinki [cit. 2022-07-01]. Dostupné z: https://siblisresearch.com/data/ftse-100-sector-weights/

Příloha A

Popisná statistika - tabulky

Sektory	Průměr [$]	Rozptyl [$]	Směr.odch. [$]	Min. [$]	Max. [$]	Modus [$]	Median [$]	Max. - Min [$]	Šikmost	Špičatost	Počet firem
Komunikační technologie	182,82	12759,88	112,96	31,20	499,80	201,50	224,70	468,60	0,67	2,94	3
Nepotřebné zboží	2148,21	3561927,86	1887,31	116,28	9786,04	250,00	2755,00	9669,76	1,16	4,12	21
Potřebné zboží	1585,94	3230070,08	1797,24	25,95	8108,00	235,00	1894,00	8082,05	1,49	4,67	8
Energie	1262,55	107607,80	328,04	101,60	1703,21	1547,00	1389,00	1601,61	-1,81	6,04	4
Finance	778,57	730982,49	854,98	21,84	7760,00	102,50	416,29	7738,16	2,56	12,77	32
Zdravotní péče	1780,93	2557454,86	1599,20	12,20	7808,00	1142,00	1462,20	7795,80	1,14	3,89	7
Průmyslová odvětví	1463,38	2164428,29	1471,20	55,18	9715,00	1060,00	1109,00	9659,82	2,34	9,04	30
Informační technologie	879,87	585541,51	765,21	183,15	3787,94	575,00	463,40	3604,79	1,96	5,96	6
Materiály	1868,44	2878054,22	1696,48	68,62	13841,01	1160,00	1767,60	13772,39	2,26	10,88	30
Realitní kanceláře	761,20	59606,96	244,15	239,20	1453,87	532,00	772,50	1214,67	0,62	2,93	3
Služby (voda, plyn atp.)	893,66	367532,23	606,24	64,58	2553,00	457,95	807,03	2488,42	0,94	2,91	5

Tabulka A.1: Popisné statistiky cen akcií podle sektorů firem spadající pod index FTSE

Sektory	Průměr [$]	Rozptyl [$]	Směr.odch. [$]	Min. [$]	Max. [$]	Modus [$]	Median [$]	Max. - Min [$]	Šikmost	Špičatost	Počet firem
Komunikační technologie	100,28	39428,28	198,57	5,73	1362,47	35,00	33,80	1356,74	3,81	18,31	27
Nepotřebné zboží	136,47	101691,49	318,89	2,99	3892,89	31,86	59,91	3889,90	5,86	45,23	60
Potřebné zboží	63,53	1455,59	38,15	6,22	305,21	28,90	59,19	298,98	1,61	7,20	31
Energie	61,59	1140,59	33,77	6,73	233,07	34,97	56,35	226,34	1,07	4,84	21
Finance	67,49	3467,90	58,89	3,02	593,26	51,81	51,27	590,24	2,80	15,17	65
Zdravotní péče	91,39	6873,85	82,91	4,47	872,45	58,00	66,16	867,98	2,76	14,68	64
Průmyslová odvětví	80,08	4085,02	63,91	3,63	597,78	38,00	72,27	594,15	1,98	8,59	74
Informační technologie	59,53	2376,54	48,75	0,70	331,20	24,70	50,95	330,50	1,82	6,99	74
Materiály	67,28	1779,75	42,19	3,74	278,71	11,71	64,24	274,97	1,24	4,86	28
Realitní kanceláře	82,70	4623,78	68,00	9,70	583,02	66,50	66,04	573,32	2,40	12,05	30
Služby (voda, plyn atp.)	46,57	575,08	23,98	5,60	152,03	32,50	42,86	146,43	0,79	3,35	29

Tabulka A.2: Popisné statistiky cen akcií podle sektorů firem spadající pod index S&P

Sektory	Průměr [%]	Rozptyl [%]	Směr.odch. [%]	Min. [%]	Max. [%]	Median [%]	Max. - Min [%]	Šikmost [%]	Špičatost [%]	Počet firem
Komunikační technologie	0,03	2,47	1,57	-20,79	25,95	0	46,74	0,83	28,53	3
Nepotřebné zboží	0,03	2,63	1,62	-22,54	27,39	0,04	49,93	-0,43	15,67	21
Potřebné zboží	0,02	2,03	1,42	-18,55	20,80	0,01	39,35	0,09	14,77	8
Energie	-0,01	3,51	1,87	-26,45	16,23	0,05	42,68	-1,73	29,47	4
Finance	0,02	3,37	1,84	-66,22	25,73	0	91,94	-1,25	51,62	32
Zdravotní péče	0,03	2,65	1,63	-32,40	36,83	0,04	69,23	-0,23	59,30	7
Průmyslová odvětví	0,04	2,85	1,69	-26,72	26,24	0,04	52,96	-0,22	13,40	30
Informační technologie	0,08	3,95	1,99	-46,35	40,87	0,08	87,23	-1,54	102,18	6
Materiály	0,01	4,49	2,12	-29,42	26,63	0	56,05	-0,06	11,78	30
Realitní kanceláře	0,02	1,89	1,38	-19,54	8,68	0,05	28,22	-0,71	15,96	3
Služby (voda, plyn atp.)	0,02	1,77	1,33	-19,00	18,45	0,05	37,45	-0,54	20,51	5

Tabulka A.3: Popisné statistiky výnosů podle sektorů firem spadající pod index FTSE

Sektory	Průměr [%]	Rozptyl [%]	Směr.odch. [%]	Min. [%]	Max. [%]	Median [%]	Max. - Min [%]	Šikmost [%]	Špičatost [%]	Počet firem
Komunikační technologie	0,07	3,76	1,94	-36,59	42,22	0,06	78,81	0,33	26,95	27
Nepotřebné zboží	0,08	3,73	1,93	-51,16	30,60	0,07	81,76	-0,05	20,72	60
Potřebné zboží	0,05	1,71	1,31	-34,13	37,23	0,06	71,36	0,06	38,27	31
Energie	0,03	4,21	2,05	-35,42	34,39	0,02	69,81	0,09	13,50	21
Finance	0,06	2,58	1,60	-26,83	18,74	0,07	45,57	-0,16	9,47	65
Zdravotní péče	0,08	3,13	1,77	-32,65	61,91	0,08	94,56	0,63	45,17	64
Průmyslová odvětví	0,07	2,80	1,67	-28,50	37,04	0,07	65,54	0,01	13,85	74
Informační technologie	0,09	4,06	2,01	-37,37	52,29	0,09	89,66	0,41	23,47	74
Materiály	0,05	3,25	1,80	-20,33	28,66	0,06	48,99	0,04	10,72	28
Realitní kanceláře	0,05	2,11	1,45	-61,89	20,05	0,08	81,95	-1,33	56,43	30
Služby (voda, plyn atp.)	0,04	1,35	1,16	-17,96	29,39	0,06	47,34	-0,13	15,57	29

Tabulka A.4: Popisné statistiky výnosů podle sektorů firem spadající pod index S&P

Příloha B

CD

K práci je přiložený CD disk. Ten obsahuje:

- *adaboost_ftse.m* - metoda pro vytvoření modelu využívající Adaboost pro data indexu FTSE
- *adaboost_sap.m* - metoda pro vytvoření modelu využívající Adaboost pro data indexu S&P
- *bag_ftse.m* - metoda pro vytvoření modelu využívající Bag pro data indexu FTSE
- *bag_sap.m* - metoda pro vytvoření modelu využívající Bag pro data indexu S&P
- *coarse_tree_ftse.m* - metoda pro vytvoření modelu Coarse tree pro data indexu FTSE
- *coarse_tree_sap.m* - metoda pro vytvoření modelu Coarse tree pro data indexu S&P
- *fine_tree_ftse.m* - metoda pro vytvoření modelu Fine tree pro data indexu FTSE
- *fine_tree_sap.m* - metoda pro vytvoření modelu Fine tree pro data indexu S&P
- *medium_tree_ftse.m* - metoda pro vytvoření modelu Medium tree pro data indexu FTSE
- *medium_tree_sap.m* - metoda pro vytvoření modelu Medium tree pro data indexu S&P
- *opt_tree_ftse.m* - metoda pro vytvoření modelu získaného z optimalizace pro data indexu FTSE
- *opt_tree_sap.m* - metoda pro vytvoření modelu získaného z optimalizace pro data indexu S&P

- *rusboost_ftse.m* - metoda pro vytvoření modelu využívající RUSBoost pro data indexu FTSE
- *uprava_csv.m* - funkce využitá pro úpravu dat indexu FTSE
- *uprava_sap.m* - funkce využitá pro úpravu dat indexu S&P
- *FTSE_data.csv* - data firem a indexu FTSE
- *S&P_data.csv* - data firem a indexu S&P
- *vystup_ftse.csv* - výstupy modelů pro obchodováni u indexu FTSE
- *vystup_sap.csv* - výstupy modelů pro obchodováni u indexu S&P
- *bakalarska_prace.pdf* - bakalářská práce ve formátu pdf