UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

AUTOMATIZOVANÝ MODEL OCEŇOVANIA FIRIEM

DIPLOMOVÁ PRÁCA

2025 Bc. MICHAL DRGOŇ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



AUTOMATIZOVANÝ MODEL OCEŇOVANIA FIRIEM

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1113 Matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	doc. RNDr. Zuzana Chladná, Dr.

Bratislava 2025 **Bc. MICHAL DRGOŇ**





Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Michal Drgoň
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Automatizovaný model oceňovania firiem
Automated valuation model of the firms

Anotácia: Automatizované modely oceňovania (AVM) predstavujú plne automatizovaný výpočet odhadu súčasnej ceny firiem na základe verejne dostupných dát s využitím matematických a štatistických modelov a metód strojového učenia. V súčasnosti sa modely AVM využívajú najmä v oblasti oceňovania nehnuteľností. Pri oceňovaní firiem je tento prístup zatiaľ len málo rozšírený: v porovnaní s oceňovaním nehnuteľností predstavuje ocenenie firmy rádovo komplexnejší problém. Cieľom konštrukcie AVM v tejto oblasti je poskytnúť rýchly a spoľahlivý odhad ceny zvolenej firmy. Na rozdiel od tradičných oceňovacích techník AMV modely majú za cieľ odstrániť z oceňovacieho procesu subjektívny pohľad analytika, tj. pri odhade ceny využiť len dostupné verejné dáta.

Vedúci: doc. RNDr. Zuzana Chladná, Dr.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: doc. Mgr. Igor Melicherčík, PhD.

Dátum zadania: 11.01.2024

Dátum schválenia: 25.01.2024 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Rád by som najprv vyjadril úprimné poďakovanie mojej vedúcej diplomovej práce, doc. RNDr. Zuzane Chladnej, Dr., za jej odborné vedenie, cenné rady a ochotu vždy poskytnúť spätnú väzbu. Veľmi si vážim jej ľudský a chápavý prístup, ktorý významne prispel k úspešnému dokončeniu tejto práce.

Touto cestou by som zároveň rád vyjadril svoju úprimnú vďaku všetkým, ktorí ma podporovali pri písaní tejto diplomovej práce. Najväčšie poďakovanie patrí mojej manželke Adrianke za jej nekonečnú trpezlivosť, podporu a porozumenie počas celého štúdia. Nakoniec by som chcel poďakovať celej svojej rodine, ktorá mi vytvorila vhodné podmienky na prácu a poskytla mi potrebné zázemie. Vaša podpora pre mňa znamená nesmierne veľa. Ďakujem vám všetkým.

Abstrakt v štátnom jazyku

DRGOŇ, Michal: Automatizovaný model oceňovania firiem [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; vedúca práce: doc. RNDr. Zuzana Chladná, Dr., Bratislava, 2025, 90 s.

Cieľom diplomovej práce je navrhnúť automatizovaný model oceňovania verejne obchodovaných spoločností pomocou metód strojového učenia. Zameriavame sa na predikciu vybraných relatívnych násobkov na základe účtovných údajov firiem z indexu S&P 500. V práci porovnávame viaceré regresné modely vrátane lineárnej regresie, *support vector regression* (SVR) a XGBoost regresie, pričom osobitná pozornosť je venovaná porovnaniu ich presnosti a interpretácii výstupov z XGBoost. Získané predikcie sú následne využité v investičnej stratégii typu LONG-SHORT, ktorej výkonnosť je porovnaná s tradičnými prístupmi. Práca ilustruje potenciál využitia automatizovaných modelov pri hodnotení a výbere akcií založenom na fundamentálnych dátach.

Kľúčové slová: automatizované oceňovanie, strojové učenie, XGBoost, teória portfólia, finančné modelovanie

Abstract

DRGOŇ, Michal: Automated valuation model of the firms [Master Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: doc. RNDr. Zuzana Chladná, Dr., Bratislava, 2025, 90 p.

The aim of this thesis is to design an automated model for valuing publicly traded companies using machine learning methods. We focus on predicting selected relative valuation multiples based on accounting data of firms from the S&P 500 index. Several regression models are compared, including linear regression, support vector regression (SVR), and XGBoost regression, with particular emphasis on evaluating predictive accuracy and interpreting the outputs of the XGBoost model. The resulting predictions are then used in a LONG–SHORT investment strategy, which is analyzed in terms of performance and compared to traditional approaches. The thesis illustrates the potential of automated models for equity valuation and selection based on fundamental data.

Keywords: automated valuation, machine learning, XGBoost, portfolio theory, financial modeling

Obsah

Úvod **8**

1 Prehľad oceňovacích prístupov **10**

1.1 Modely diskontovaných peňažných tokov (DCF) 10

1.1.1 Podstata DCF oceňovania 12

1.1.2 Gordonov rastový model 13

1.1.3 Model diskontovaných peňažných tokov pre akcionárov 14

1.1.4 Model diskontovaných voľných peňažných tokov pre firmu 14

1.1.5 Model istotného ekvivalentu (CEQ) 15

1.1.6 Modely nadmerných výnosov 16

1.1.7 Model upravenej súčasnej hodnoty (APV) 18

1.2 Relatívne oceňovanie 18

1.3 Prepojenie DCF prístupov a relatívneho oceňovania 20

1.4 Zhrnutie oceňovacích prístupov 21

2 Štatistické metódy pri oceňovaní firiem **23**

2.1 Lineárny regresný model 23

2.1.1 Špecifikácia modelu 24

2.1.2 Odhad parametrov 24

2.2 Teória Support Vector Machines 25

2.3 Extreme Gradient Boosting (XGB) 27

3 Automatizovaný model **30**

3.1 Špecifikácia dát 30

3.1.1 Obmedzenia dát 32

3.1.2 Odstraňovanie odľahlých pozorovaní a príprava dát 32

3.2 Numerické výsledky 33

3.2.1 Intepretácia výsledkov 36

4 Riadenie portfólia **47**

4.1 Naivne diverzifikované portfólio (N) 48

4.2 Markowitzovo portfólio 49

4.2.1 *Global minimum variance* portfólio (G) 49
4.2.2 Portólio optimalizované cez Sharpeov pomer (S) 50
4.3 LONG-SHORT portfólio (LS) 51

Záver **56**

Zoznam použitej literatúry **59**

Príloha A **66**

Príloha B **69**

Príloha C **76**

Úvod

Oceňovanie aktív je problém, ktorým sa ľudstvo zaoberá už od nepamäti. Aj v dobe, keď fungoval iba barter si ľudia kládli jednoduchú otázku: *koľko kusov inej komodity viem dostať za komoditu, ktorú aktuálne vlastním?*

Doba sa posunula a aktíva, ktoré vlastníme, sú dnes podstatne komplexnejšie. Centrálnou témou tejto práce je oceňovanie špecifického typu aktív – firiem. Pri určovaní ich hodnoty existujú dva hlavné prístupy: buď firmu vnímame ako zdroj budúcich peňažných tokov, ktoré diskontujeme na ich súčasnú hodnotu, alebo ako komoditu, ktorej cena by mala zodpovedať hodnote porovnateľných firiem.

Modely založené na diskontovaní peňažných tokov označujeme ako *discounted cash flow* modely (DCF). Hoci zmienka o časovej hodnote peňazí bola už v Talmude [47], základy moderného konceptu oceňovania spoločností pomocou sofistikovanejších modelov boli položené až Millerom a Modiglianim [49]. Od polovice 20. storočia sa však vývoj v oblasti oceňovania výrazne posunul a analytici v popredných investičných bankách vytvorili široké spektrum modelov — od univerzálnych prístupov použiteľných naprieč odvetviami až po vysoko špecializované riešenia prispôsobené konkrétnym situáciám ([8], [12], [10], [41]).

Ak sa na firmu pozeráme ako na komoditu, ktorú možno vymieňať za rovnaké peňažné sumy ako podobné firmy, hovoríme o relatívnom oceňovaní. I keď tento spôsob nie je tak sofistikovaný ako DCF modely, teší sa veľkej popularite, a to aj medzi samotnými analytikmi, ktorí relatívne oceňovanie používajú ako *sense check* svojich výstupov z DCF modelov. [12]

Rozmach strojového učenia a jeho uplatnenie v rôznych oblastiach (ako možno vidieť v publikáciách [18] a [16]) otvára otázku, či by mohlo nájsť uplatnenie aj pri oceňovaní firiem. Súčasné trendy naznačujú, že by to tak mohlo byť: najvýkonnejší fond na svete využíva metódy strojového učenia na oceňovanie derivátov a otváranie krátkodobých pozícií [9].

Iný prístup k oceňovaniu a investovaniu reprezentujú *value* investori, ktorých základnou tézou je oceňovanie na základe analýzy fundamentálnych (účtovných) ukazovateľov [26]. Medzi najznámejších *value* investorov patrí Warren Buffet, ktorého profesionálna kariéra ukazuje, že aj tento investičný prístup vie byť úspešný.

Táto práca sa usiluje o prepojenie týchto prístupov – konkrétne o využitie kvantitatívnych metód na analýzu fundamentálnych ukazovateľov. Hoci je uplatňovanie metód strojového učenia v rôznych oblastiach na vzostupe, v slovenskej a českej odbornej literatúre sa tejto problematike zatiaľ venovala len okrajová pozornosť, čo potvrdzuje aj skutočnosť, že sme neidentifikovali žiadnu prácu s podobným zameraním. Za zmienku však stojí niekoľko odborných článkov, ktoré sa stali dôležitým východiskom pre tento výskum. Ako prvý spomenieme článok [25], v ktorom autori využívajú balík *XGBoost*, jeden z najsilnejších algoritmov strojového učenia, na predikovanie relatívnych násobkov, pričom ide o prácu, ktorá si vďaka svojej prínosnosti vyslúžila značný počet citácií. V predkladanej práci nadväzujeme aj na autorov článkov [30] a [51], ktorí síce pracujú s odlišnými dátami, no ich prístupy boli rovnako inšpiratívne. Práca je rozdelená do štyroch kapitol. Prvá kapitola sa venuje prehľadu základných prístupov oceňovania firiem. Pri spracovaní teoretickej časti vychádzame z odbornej literatúry, ktorá zahŕňa publikácie autorov z prostredia investičného bankovníctva aj akademickej sféry. V druhej kapitole stručne predstavujeme tri regresné modely: lineárnu regresiu, *support vector regression (SVR)* a XGBoost. Výber týchto modelov pokrýva spektrum od klasickej štatistickej metódy (lineárna regresia), cez etablovaný model strojového učenia (SVR), až po moderný a výkonný algoritmus s vysokou predikčnou presnosťou (XGBoost). Tretia kapitola sa venuje dátam potrebným na konštrukciu automatizovaného oceňovacieho modelu. Následne opisujeme samotný modelovací postup a numericky porovnávame výkonnosť všetkých troch regresných prístupov. V poslednej kapitole nadväzujeme na predchádzajúce výsledky a využívame ich na konštrukciu portfóliovej investičnej stratégie. Impulzom pre zaradenie tejto časti, ktorá je nad rámec pôvodného zadania práce, bol článok [30], pričom samotný koncept stratégie sme navrhli samostatne a nezávisle. Výkonnosť navrhovaného prístupu následne porovnávame so štandardnými metódami.

1 Prehľad oceňovacích prístupov

Problematika oceňovania firiem tvorí dôležitú súčasť podnikových financií a investičného rozhodovania. Cieľom oceňovania je určiť ekonomickú hodnotu podniku alebo jeho časti na základe dostupných informácií o jeho finančnej situácii, očakávanom vývoji a trhovom prostredí. Hodnota podniku pritom nemusí byť identická s jeho trhovou cenou, ktorá môže byť ovplyvnená krátkodobými nerovnováhami, informačnými asymetriami alebo iracionálnym správaním účastníkov trhu.

V ekonomickej literatúre sa môžeme stretnúť s viacerými prístupmi k oceňovaniu. Odlišujú sa najmä v tom, aký pohľad na hodnotu podniku zohľadňujú, a aké predpoklady o budúcom vývoji prijímajú. V tejto kapitole vychádzame najmä z [8], ale informácie dopĺňame aj z [41], [11] a [12]. Uvedené zdroje poskytujú moderný prehľad oceňovacích metód používaných najväčšími investičnými bankami.

V nasledujúcich častiach jednotlivé oceňovacie prístupy predstavíme. Zameriame sa nielen na ich teoretické pozadie, ale predstavíme aj možnosti ich praktickej aplikácie.

1.1 Modely diskontovaných peňažných tokov (*DCF*)

Modely diskontovaných peňažných tokov predstavujú jadro fundamentálneho oceňovania aktív. Ich základným predpokladom je, že hodnota akéhokoľvek finančného aktíva je odvodená zo súčasnej hodnoty očakávaných budúcich peňažných tokov, ktoré dané aktívum generuje. Matematickú formuláciu tejto idey uvádzame v definícii 1.1.1

Definícia 1.1.1. Nech $t = 1, \ldots, N$ označuje časovú periódu, nech CF_t je peňažný tok generovaný aktívom v čase t. Nech r označuje zvolenú diskontnú mieru. Potom pod súčasnou hodnotou aktíva rozumieme

$$V_0 = \sum_{t=1}^{N} \frac{CF_t}{(1+r)^t}. \tag{1}$$

Prístup z definície 1.1.1 je teoreticky ukotvený v koncepte časovej hodnoty peňazí a umožňuje flexibilnú aplikáciu naprieč rôznymi typmi aktív a účelmi oceňovania. Teoretické základy pre oceňovanie firiem pomocou DCF modelov položili Miller a Modigliani [49].

Poznámka 1.1.2. Poznamenajme, že diskontná sadzba nemusí byť v čase konštantná. V takom prípade je vhodné uvažovať o časovo variabilnej diskontnej miere, ktorú označu-

jeme ako r_t. Toto zovšeobecnenie umožňuje lepšie reflektovať dynamiku rizika, najmä v situáciách s dlhým plánovacím horizontom alebo premenlivými trhovými podmienkami.

Pri pohľade na rovnicu (1) vidíme, že na riziko spojené s peňažnými tokmi sa dá nahliadnuť dvojako. Buď je riziko reflektované v diskontnej miere r alebo v samotnej veľkosti CF_t.

Všetky modely diskontovaných peňažných tokov však potrebujú odpovede na tri otázky

1. komu patria peňažné toky?
2. akým spôsobom sa menia peňažné toky v čase?
3. aké riziko je späté s peňažnými tokmi?

V kontexte oceňovania firiem, na ktoré sa v predkladanej práci sústredíme, je odpoveď na prvú otázku jednoduchá - buď patria peňažné toky akcionárom alebo akcionárom spolu s veriteľmi. Typy peňažných tokov definujeme v samotných modeloch, keďže sa pre každý model odlišujú.

V prípade, že ohodnocujeme peňažné toky očami akcionára, ľavú stranu rovnice (1) budeme špeciálne nazývať súčasná hodnota vlastného imania. Budeme ju značiť E, horným indexom označíme typ modelu, prostredníctvom ktorého bola táto hodnota zrátaná. Ak realizujeme oceňovanie peňažných tokov, ktoré prináležia všetkým veriteľom, hovoríme o súčasnej hodnote firmy, ďalej značenej ako EV, podľa horného indexu označíme typ modelu.

Odpoveď na druhú otázku je dôležitá iba v prípade, že chceme ukázať špeciálne vlastnosti týchto modelov. DCF modely vedú spravidla na tvar rovnice (1). Za určitých predpokladov možno pravú stranu stranu (1) zapísať zjednodušene. V prekladanej práci sa na tieto predpoklady nezameriavame; ak ich však niekde využijeme, náležite na to upozorníme.

Poslednou nezodpovedanou otázkou zostáva cez akú prizmu sa máme pozerať na riziko späté s peňažnými tokmi. V aplikačnej praxi je bežne využívaný prístup vyjadrenia rizika cez diskontný faktor. V závislosti od kategórie peňažného toku, buď pre akcionára alebo všetkých veriteľov, zavádzame rôzne diskontné miery. Prirodzene platí, že väčšia diskontná sadzba odráža väčšie riziko a naopak. Táto logika je čiastočne preberaná aj v článku [42].

Ak diskontujeme bezrizikové peňažné toky (to sú také, ktoré nadobudnú svoju hodnotu s istotou), je správne použiť tzv. bezrizikovú diskontnú mieru, ktorú značíme r_f. O tom, čo je vlastne bezriziková miera, sa stále vedú diskusie. Ďalšou podstatnou mierou je tzv. náklad na dlh, ktorý značíme r_d. Ďalšie podrobnosti o význame ako aj o spôsobe odhadu týchto diskontných mier možno nájsť v [14].

V prípade peňažných tokov pre akcionárov diskontujeme pomocou nákladu na vlastné imanie, značíme r_E. Možno ho odhadnúť pomocou *capital asset pricing* modelu (CAPM) [55] ako

$$r_E = r_f + \beta(E[R_m] - r_f), \tag{2}$$

kde $E[R_m]$ je očakávaný výnos trhu a β je koeficient systematického rizika firmy .

Diskont peňažných tokov určených akcionárom aj veriteľom vykonávame váženým nákladom na kapitál, tzv. $WACC$ a vypočítame ho ako

$$WACC = \frac{E_{opt}}{Debt_{opt} + E_{opt}} r_E + \frac{Debt_{opt}}{Debt_{opt} + E_{opt}} r_D(1 - \text{Tax}), \tag{3}$$

kde $Debt_{opt}$ je trhová hodnota cudzieho kapitálu a Tax je efektívna daňová sadzba.

Poznámka 1.1.3. V rovnici (3) si môžeme všimnúť, že veličiny E a Debt na pravej strane nesú index $_{opt}$. Hodnota vlastného imania E je pritom výsledkom samotného ocenenia, zatiaľ čo $WACC$ predstavuje vstup do tohto procesu. Z toho vyplýva, že medzi týmito veličinami existuje implicitný vzťah. Kapitálová štruktúra sa preto v praxi často vyjadruje prostredníctvom tzv. optimálnej štruktúry, o ktorej sa predpokladá, že je nemenná počas celého obdobia, preto používame index $_{opt}$.

1.1.1 Podstata DCF oceňovania

Za najjednoduchší model diskontovaných peňažných tokov je považovaný Gordonov rastový model. Tento model však pracuje s dividendami, čo prirodzene znižuje jeho možnosť využitia v aplikačnej praxi. Nie všetky firmy totiž vyplácajú dividendy.

Za univerzálnejšie sú považované diskontné modely, ktoré možno v závislosti od typu oceňovaného peňažného toku systematizovať do dvoch hlavných kategórií: modely zamerané na hodnotu pre akcionárov (tzv. equity DCF) a modely orientované na hodnotu

celej firmy (tzv. entity DCF). V odbornej literatúre sa tieto prístupy často označujú aj ako model založený na diskontovaní voľných peňažných tokov pre akcionárov (DCF FCFE) a model založený na diskontovaní voľných peňažných tokov pre firmu (DCF FCFF). Okrem nich sa v aplikačnej praxi používajú modely istotných ekvivalentov (*certainty equivalent*, CEQ), modely nadmerných výnosov (*excess return model*, ERM) a model upravenej súčasnej hodnoty (*adjusted present value*, APV).

1.1.2 Gordonov rastový model

Tento model vychádza z predpokladu, že pri kúpe určitého typu akcií dostáva investor v pravidelných intervaloch t dividendu DD_t. Okrem dividendy očakáva ešte inkasovanie peňažného toku pri predaji akcie. Keďže cena samotná sa odvíja od očakávanej výšky dividendy (aj za horizontom predaja akcie), dostávame, že súčasná hodnota vlastného imania spojená s vlastníctvom jednej akcie sa dá vyjadriť ako

$$E^G = \sum_{t=1}^{\infty} \frac{DD_t\,(1+g)}{(1+r_E)^t}, \tag{4}$$

kde DD_t je výška dividendy a $g > 0$ je rastová konštanta. V základnom Gordonovom rastovom modeli sa predpokladá, že $DD_t = DD_1$ pre všetky t.

Využívajúc logický predpoklad $r_E > g$ a vykonajúc triviálne úpravy, geometrický rad z pravej strany rovnice (4) možno upraviť na

$$E^G = \frac{DD_1}{r_E - g}. \tag{5}$$

Poznámka 1.1.4. Pri konštrukcii modelu sme využili jeden z valuačných predpokladov, a to konkrétne *going concern*. Teda, že firma funguje a vypláca dividendy na nekonečnom časovom horizonte.

Zo zápisu modelu (4) je vidno, že model má aj svoje slabiny. Predpokladá totiž konštantný rast dividend a nemennú diskontnú sadzbu. V publikácii [10] sú v piatej kapitole popísané aj rozšírenia Gordonovho modelu. Tento model sme do výberu zaradili preto, lebo poskytuje vhodný odrazový mostík pre iné typy DCF modelov.

1.1.3 Model diskontovaných peňažných tokov pre akcionárov

V tomto modeli sa diskontujú peňažné toky, ktoré sú k dispozícii vlastníkom akcií. Tie zohľadňujú čistý zisk po zdanení, nepeňažné nákladové položky, zmenu v pracovnom kapitáli a zmenu úročených záväzkov.

Z formálneho hľadiska peňažný tok pre akcionárov v čase t definujeme ako

$$FCFE_t = NI_t + D_t - CAPEX_t - \Delta WC_t + \Delta Debt_t, \tag{6}$$

kde NI_t je čistý zisk, D_t odpisy, $CAPEX_t$ kapitálové výdavky, ΔWC_t zmena pracovného kapitálu a $\Delta Debt_t$ čistý prírastok dlhu v časovej perióde t. Výsledná hodnota vlastného imania sa získava ako

$$E^{FCFE} = \sum_{t=1}^{N} \frac{FCFE_t}{(1 + r_E)^t}. \tag{7}$$

1.1.4 Model diskontovaných voľných peňažných tokov pre firmu

Alternatívne možno diskontovať finančné toky, ktoré sú dostupné všetkým investorom: vlastníkom aj veriteľom. Peňažný tok pre firmu ($FCFF$) je definovaný nasledovne

$$FCFF_t = EBIT_t(1 - \text{Tax}) + D_t - CAPEX_t - \Delta WC_t, \tag{8}$$

kde $EBIT_t$ je zisk pred úrokmi a zdanením a Tax efektívna daňová sadzba, t opäť označuje časovú periódu. Peňažné toky (8) sa diskontujú pomocou $WACC$, z čoho sa dá odvodiť hodnota firmy

$$EV^{FCFF} = \sum_{t=1}^{N} \frac{FCFF_t}{(1 + WACC)^t}. \tag{9}$$

Hodnota vlastného imania E^{FCFF} sa potom získa odpočítaním D_0 a pripočítaním C_0 od EV^{FCFF}, teda

$$E^{FCFF} = EV^{FCFF} - Debt_0 + Cash_0. \tag{10}$$

Tak ako je uvedené v prílohe C v [41], za použitia konzistentných predpokladov, model založený na voľných peňažných tokoch pre akcionárov a model založený na voľných peňažných tokoch pre firmu dávajú rovnakú hodnotu vlastného imania, čo sa formálne dá zapísať

$$E^{FCFF} = E^{FCFE}. \tag{11}$$

1.1.5 Model istotného ekvivalentu (CEQ)

Model istotného ekvivalentu je skôr teoretickým konceptom, keďže jeho využitie v praxi je, ako vysvetlíme neskôr, značne problematické. Napriek tomu poukazuje na zásadný fakt - v ostatných modeloch sa do úvahy vôbec neberie subjektívne riziko vnímané investorom. Model nadväzuje na koncepty práce [4], v ktorej bol prezentovaný rámec pre teóriu užitočnosti. Značný priestor teoretickému odvodeniu tohto modelu je venovaný v kapitole 43 v knihe [8]. Náznak praktickej aplikácie modelu istotných ekvivalentov možno nájsť v [13].

K predstaveniu modelu istotných ekvivalentov potrebujeme ešte definovať špeciálny typ peňažného toku (viď. definícia 1.1.5).

Definícia 1.1.5. Nech $t = 1, \ldots, N$ je časová perióda. Nech $s = 1, \ldots, S$ index označujúci možný stav peňažných tokov. Potom pod peňažným tokom určeným pre akcionára s rôznymi možnými budúcimi stavmi rozumieme $FCFE_{t,s}$.

Nech U je funkcia užitočnosti akcionára. Potom pod istotným ekvivalentom peňažného toku pre akcionára rozumieme

$$CEQ_t^E = U^{-1}\left(E\left(U\left(FCFE_{t,s}\right)\right)\right). \tag{12}$$

Teraz nám už nič nebráni v tom, aby sme spočítali hodnotu E^{CEQ}. Máme

$$E^{CEQ} = \sum_{t=1}^{N} \frac{CEQ_t^E}{\left(1 + r_f\right)^t}. \tag{13}$$

Poznámka 1.1.6. Z definície 1.1.5 je zrejmé, že podobným spôsobom ako v (13) by sme vedeli odvodiť aj peňažné toky pre firmu s rôznymi možnými budúcimi stavmi $FCFF_{t,s}$. Potom by sme vedeli spočítať aj EV^{CEQ}.

Poznámka 1.1.7. Všimnime si, že v (13) diskontujeme bezrizikovovou mierou r_f. Dôvodom je, že riziko je subjektívne už zohľadnené funkciou U každého investora.

Poznámka (1.1.7) ukazuje najväčšiu slabinu modelu istotných ekvivalentov. Pre každého investora je hodnota E^{CEQ} rozdielna.

Tak ako je naznačené v [13] model môže mať svoje uplatnenie. Napríklad, niektorí analytici pri oceňovaní aktív znížia peňažné toky predelením pomocou tzv. *haircut*. CEQ_t^E

potom dostaneme ako

$$CEQ_t^E = \frac{FCFE_t}{\phi_t}, \tag{14}$$

kde $\phi_t > 1$ je *haircut*. *Haircut* však analytici volia subjektívne.

Dosadením pravej strany (14) do vzťahu (13) dostávame súčasnú hodnotu vlastného imania

$$E^{CEQ} = \sum_{t=1}^{N} \frac{\frac{FCFE_t}{\phi_t}}{(1+r_f)^t}. \tag{15}$$

Napriek značným slabinám existujú aj autori, ktorí model istotných ekvivalentov obhajujú. Ich snaha spočíva v zobjektívnení hľadania istotných ekvivalentov pomocou sofistikovaných štatistických metód. Presvedčiť sa o tom možno v [19]. Ich primárnym argumentom je, že odhady nákladov na vlastné imanie r_E generované pomocou CAPM modelu neodrážajú ekonomickú realitu, a to najmä v prípade firiem, ktoré nie sú obchodované.

1.1.6 Modely nadmerných výnosov

Ďalšiu alternatívu k tradičným DCF modelom predstavujú modely založené na tzv. *nadmerných výnosoch*. Vychádzajú z myšlienky, že hodnota podniku je daná súčtom účtovnej hodnoty investovaného kapitálu a súčasnej hodnoty nadmerných výnosov. Nadmerný výnos RI_t definujeme ako príjmy presahujúce absolútny náklad na kapitál požadovaný akcionárom. Vyjadrime ho ako

$$NI_t = RI_t + r_E \times BV_t, \tag{16}$$

kde BV_t je účtovná hodnota vlastného imania v čase t a NI_t je čistý zisk v čase t. Vzťah (16) nám umožňuje dopočítať RI_t zo znalosti NI_t, BV_t a r_E.

Pre rozvinutie ďalších myšlienok potrebujeme definovať rentabilitu vlastného imania. Podľa [41] definujeme rentabilitu vlastného imania v čase t ako

$$ROE_t = \frac{NI_t}{BV_t}. \tag{17}$$

Dá sa nahliadnuť, že RI_t vieme zrátať ako kombináciu (17) a (16) ako

$$RI_t = (ROE_t - r_E) \times BV_t. \tag{18}$$

Poznámka 1.1.8. Na základe vzťahu (18) vidíme, že akcionári by mali zvážiť iba tie investície, ktorých návratnosť presahuje r_E.

Modely nadmerných výnosov rozčleňujú hodnotu firmy na dve zložky: kapitál investovaný k dnešnému dňu a súčasnú hodnotu budúcich nadmerných výnosov z existujúcich a budúcich projektov.

V modeloch nadmerných výnosov hodnotu vlastného imania vyjadríme nasledovne

$$E^{ER} = BV_0 + \sum_{t=1}^{T} \frac{RI_t}{(1+r_E)^t}. \tag{19}$$

Na vyjadrenie hodnoty firmy EV^{ER}, musíme použiť účtovné ukazovatele, ktoré prináležia ako akcionárom, tak aj veriteľom. Pri ich vyjadrení vychádzame z [41]. Ekvivalentom čistého zisku pre firmu je tzv. prevádzkový zisk po zdanení ($NOPLAT_t$, v angličtine *net operating profit less adjusted taxes*). $NOPLAT_t$ predstavuje súčet výnosu požadovaného akcionármi a veriteľmi spolu s dodatočnou hodnotou vytvorenou nad rámec ich očakávaní, čo vyjdríme ako

$$NOPLAT_t = EVA_t + WACC \times IC_t, \tag{20}$$

kde EVA_t je nadmerný výnos firmy (*economic value added*) v čase t a IC_t je investovaný kapitál akcionármi aj veriteľmi v čase t. Ďalej rentabilitou investovaného kapitálu v čase t rozumieme

$$ROIC_t = \frac{NOPLAT_t}{IC_t}. \tag{21}$$

Z rovníc (20) a (21) vyjadríme EVA_t nasledovne

$$EVA_t = (ROIC_t - WACC) \times IC_t. \tag{22}$$

V modeloch nadmerných výnosov vyjadrujeme hodnotu firmy ako

$$EV^{ERM} = IC_0 + \sum_{t=1}^{T} \frac{EVA_t}{(1+WACC)^t}. \tag{23}$$

Tento typ ocenenia je vhodný najmä pre finančné inštitúcie (najmä banky a poisťovne). Účtovné výkazy týchto inštitúcií zvyčajne poskytujú hodnoverný obraz o stave aktív a pasív. Ak to dáme do kontrastu napríklad so stavebnými firmami, ktoré majú typicky veľké množstvo strojov či iného hmotného majetku - ten často v účtovníctve nebýva správne precenený, účtovná čiastka býva často skreslená aj odpisovaním majetku [12].

Výhodou modelu nadmerných výnosov je to, že pri oceňovaní nie je nutné odhadovať peňažné toky. Nevýhodou je veľká citlivosť na kvalitu účtovníctva.

1.1.7 Model upravenej súčasnej hodnoty (APV)

Ak sa poriadne pozrieme na rovnicu (3), zistíme, že pri výpočte $WACC$ využívame niekoľko veľmi obmedzujúcich predpokladov: ako príklad môžeme uviesť konštantnosť kapitálovej štruktúry. V prípadoch, keď oceňujeme firmu, ktorej kapitálová štruktúra sa bude s veľkou pravdepodobnosťou významne meniť, je vhodné použiť model upravenej súčasnej hodnoty.

Podľa APV modelu vypočítame hodnotu firmy nasledujúcim spôsobom

$$EV^{APV} = V^{\text{Nezadĺžená}} + PV(\text{Tax shield}) - PV(\text{Finančné ťažkosti}), \tag{24}$$

kde $V^{Nezadĺžená}$ predstavuje hodnotu firmy financovanej výlučne vlastným kapitálom, $PV(\text{Tax shield})$ je súčasná hodnota daňových úspor z úrokových nákladov, a $PV(\text{Finančné ťažkosti})$ sú očakávané náklady na finančné ťažkosti, ktoré pramenia z vysokej rizikovosti zadlžovania sa.

Podľa [41] hodnota firmy bez dlhu ráta ako diskontovaný súčet voľných peňažných tokov pre firmu (FCFF) pri použití nezadlžených nákladov vlastného kapitálu

$$V^{\text{Nezadĺžená}} = \sum_{t=1}^{T} \frac{FCFF_t}{(1+r_0)^t}, \tag{25}$$

kde r_0 je náklad vlastného kapitálu pre firmu bez dlhu.

Daňový štít možno v základnej verzii APV odhadnúť ako

$$PV(\text{Tax Shield}) = \sum_{t=1}^{T} \frac{\text{Tax} \cdot r_D \cdot Debt_t}{(1+r_D)^t}. \tag{26}$$

Výhodou prístupu je, že samostatne analyzuje hodnotu samotného projektu a prínosy či náklady súvisiace s jeho financovaním, čím poskytuje väčšiu flexibilitu v modelovaní zmien kapitálovej štruktúry v čase.

Nevýhodou je nutnosť odhadovať náklad na vlastné imanie pre firmu bez dlhu r_0. Spôsoby, akým sa táto veličina dá odhadnúť, sú detailne popísané v [41].

1.2 Relatívne oceňovanie

Relatívne oceňovanie predstavuje jednu z najpoužívanejších metód oceňovania aktív v praxi [12]. Jej základnou filozofiou je oceňovať aktívum nie na základe jeho vlastných peňažných tokov a rizika s ním spojeným, ale na základe cien podobných aktív, ktoré

sa už na trhu obchodujú. Problematika relatívneho oceňovania je podrobne spracovaná napr. v [12].

Poznámka 1.2.1. Z uvedeného vyplýva, že ak sa účastníci na trhu nemýlia a všetky parametre v DCF modeloch by boli odhadnuté správne, relatívne ocenenie a DCF ocenenie by k sebe mali konvergovať [12].

Na rozdiel od DCF oceňovania, pri ktorom je nutné odhadnúť obrovské množstvo vstupov, pri relatívnom oceňovaní je nutné urobiť iba tri kroky.

Prvým krokom je nájsť skupinu podobných firiem (po anglicky *peer group*). V literatúre je popísaných množstvo prístupov, ako túto skupinu hľadať. Medzi časté spôsoby patrí podľa odvetvia, geografie či biznis modelu [41]. Medzi nápaditejšie prístupy patria prístupy prezentované v [31] a [45].

V druhom kroku zrátame vhodný násobok, pomocou ktorého firmy medzi sebou porovnáme. Pomerový ukazovateľ (násobok) zrátame ako

$$\text{Násobok} = \frac{\text{To, čo za firmu platíme}}{\text{To, čo za firmu dostaneme}}. \tag{27}$$

Násobok na ľavej strane (27) zrátame pre každú firmu zo skupiny porovnateľných firiem. Z týchto násobkov vyrátame priemer alebo medián [12]. Proces zrátavania násobkov sa nazýva štandardizácia. Štandardizácia je potrebná na odstránenie rozdielov v charakteristikách porovnávaných firiem, ako je napríklad veľkosť či cenová úroveň akcií, aby bolo porovnanie relevantné. V treťom kroku rozhodneme, či je násobok firmy, ktorú ohodnocujeme, väčší alebo menší ako výsledok, ktorý sme dostali v druhom kroku: porovnaním zistíme, či je firma nadhodnotená alebo podhodnotená.

Hoci je postup priamočiary, je nutné vyvarovať sa chybám, ktoré môžu pri výpočte vzniknúť. V prvom rade je nutné zabezpečiť konzistentnosť násobkov z rovnice (27). Ak napríklad v menovateli počítame s hodnotou, ktorá súvisí s celou firmou (nie iba s akcionármi), do menovateľa by mala ísť rovnako hodnota, ktorá súvisí s celou firmou. Medzi najčastejšie používané násobky patrí $\frac{mEV_t}{EBITDA_t}$, kde $EBITDA_t$ označuje zisk firmy pred zdanením, odpismi a úrokmi v čase t. Označenie mEV_t sa používa na označenie hodnoty firmy v čase t. Táto hodnota sa však líši od hodnoty firmy v DCF modeloch a rátame ju ako

$$mEV_t = MC_t + Debt_t - Cash_t, \tag{28}$$

kde MC_t označuje trhovú kapitalizáciu firmy v čase t [12]. Okrem toho je často používaným násobkom aj $\frac{P}{E}$, ktorý rátame ako

$$\frac{P}{E} = \frac{\text{Cena akcie v čase } t}{NI_t}. \tag{29}$$

Pri rátaní priemeru a mediánu násobkov často nevychádzame iba z malej skupiny firiem. Ak to množina porovnateľných firiem dovoľuje, mali by sme pre násobok zostaviť aj empirickú distribúciu. Takáto štatistická analýza nám môže poskytnúť množstvo cenných informácií: napríklad či sú dáta šikmé alebo či existuje veľké množstvo odľahlých pozorovaní. Treba mať na pamäti, že hodnoty násobkov majú tendenciu sa v čase meniť. Ako uvádza autor v [12], za najpravdepodobnejšie príčiny môžno považovať tri fakty - meniaciu sa úroveň úrokových sadzieb, všeobecný rast ekonomickej úrovne a predovšetkým vnímanie rizika. Ak teda porovnávame historické hodnoty relatívnych násobkov pre nejakú firmu, mali by sme do úvahy brať aj časový aspekt.

1.3 Prepojenie DCF prístupov a relatívneho oceňovania

Vo všeobecnosti je možné každý jeden relatívny násobok previesť na výraz, v ktorom vystupujú iné ukazovatele z účtovníctva (využívajúc účtovné identity). V tejto časti uvedieme tri ukazovatele, ktoré sú predstavené aj odvodené v [25]. Tieto tri ukazovatele sme vybrali práve preto, lebo ich budeme používať v ďalších častiach práce pri konštrukcii automatizovaného modelu oceňovania firiem.

Prvým ukazovateľom je pomer trhovej kapitalizácie a účtovnej hodnoty vlastného imania BV_t

$$m2b_t = \frac{MC_t}{BV_t}. \tag{30}$$

Úpravami môžeme podľa [25] $m2b_t$ vyjadriť ako

$$m2b_t = ROE_t \left[1 - \theta_t^E (1 - \delta)\right] \frac{1+g}{r_E - g}, \tag{31}$$

kde $\theta_t^E = \frac{CAPEX_t - D_t + \Delta WC_t}{NI_t}$ a $\delta = \frac{D_{opt}}{D_{opt}+E_{opt}}$. Predpokladá sa, že δ je v čase konštantná a optimálna [25].

Druhým násobkom, ktorý uvedieme, je podiel hodnoty firmy na trhu a jej celkových aktív, t.j.

$$ev2a_t = \frac{mEV_t}{A_t}, \tag{32}$$

kde symbol A_t reprezentuje celkové aktíva v čase p. Podľa [25] sa (32) dá prepísať ako

$$ev2a_t = ROA_t \left(1 - \theta_t^A\right) \frac{1 + g_f}{WACC - g_f}, \tag{33}$$

kde ROA_t vypočítame ako $\frac{NI_t}{A_t}$, g_f reprezentuje rast $FCFF_t$ a hodnotu θ_t^A vypočítame ako $\frac{CAPEX_t - D_t + \Delta WC_t}{EBIT_t(1-\text{Tax})}$.

Posledným relatívnym násobkom je podiel hodnoty firmy na trhu a tržieb

$$ev2s_t = \frac{mEV_t}{S_t}, \tag{34}$$

kde S_t sú tržby v čase t.

Je zrejmé, že $ev2s_t$ získame prenásobením (33) $\frac{A_t}{S_t}$. Špeciálne teda vieme vzťah (34) vyjadriť ako

$$ev2s_t = \frac{ROA_t}{ATR_t} \left(1 - \theta_t^A\right) \frac{1 + g_f}{WACC - g_f}, \tag{35}$$

kde ATR_t je obrat aktív v čase t, ktorý možno zrátať ako $\frac{S_t}{A_t}$. Pravé strany rovníc (31), (33) a (35) nám umožňujú analyzovať čiastkové efekty účtovných ukazovateľov na násobky na ľavých stranách uvedených rovníc. Vo všetkých troch prípadoch je zrejmé, že s rastom ROE_t, ROA_t, NI_t a $EBIT_t$, *ceteris paribus*, rastú aj násobky na ľavých stranách (31), (33) a (35). Pozitívny vplyv má aj rast parametrov g a g_f, čo je v súlade s ekonomickou logikou. So zmenšujúcou sa daňovou sadzbou Tax relatívne násobky taktiež rastú. Negatívny vplyv majú investície do majetku ($CAPEX_t$ a ΔWC_t). Biznisovú logiku podporuje aj to, že zvýšenie nákladov na kapitál $WACC$, resp. r_E povedie k nižším násobkom. To je presne v súlade s DCF modelmi [25].

Poznámka 1.3.1. Za zmienku stojí aj fakt, že niektoré účtovné položky ovplyvňujú násobky z rovníc (31), (33) a (35) nelineárne.

1.4 Zhrnutie oceňovacích prístupov

Ako sme ukázali v predchádzajúcich častiach, DCF modely sú flexibilnejšie: vieme pomocou nich oceniť rôzne typy firiem. Na druhú stranu, proces oceňovania pomocou DCF modelov vyžaduje odhad budúcich peňažných tokov, spravidla na dlhé časové obdobie. Odhad budúcich peňažných tokov ovplyvňuje množstvo parametrov, akými sú napríklad

marže alebo $CAPEX$. Navyše, DCF prístup vyžaduje odhadovať diskontnú sadzbu. Dohľadávať, nastavovať a odhadovať spomínané vstupy pre obrovské množstvo firiem je v rozumnej časovej škále nemožné.

Pri zostavovaní automatizovaného modelu v ďalších častiach práce budeme preto vychádzať z princípov relatívneho oceňovania. Relatívne oceňovanie totiž využíva štandardizované a široko dostupné údaje, ako sú účtovné výkazy a trhové ceny. Ďalšou motiváciou pre voľbu relatívneho oceňovania je súčasné smerovanie akademického výskumu, ktorý čoraz častejšie prepája metódy strojového učenia s oceňovaním firiem [17] [3] [28], riadením rizika [21] či riadením portfólia [30]. Domnievame sa, že aplikácia moderných ML techník v tejto oblasti predstavuje perspektívny a inovatívny prístup. Hoci v zahraničnej literatúre ide o čoraz viac rozvíjanú tému, v slovenských a českých akademických publikáciách sme podobné príspevky nenašli.

2 Štatistické metódy pri oceňovaní firiem

V tejto kapitole stručne predstavíme teoretické východiská troch štatistických prístupov, ktoré následne využijeme v procese automatizovaného oceňovania firiem. Spomedzi širokého spektra dostupných štatistických techník sme sa vzhľadom na povahu skúmaného problému rozhodli implementovať nasledujúce techniky: lineárnu regresiu, *support vector machines* (SVM) (resp. *support vector regression* (SVR)) a *extreme gradient boosting* (XGB).

Lineárna regresia patrí medzi kľúčové metódy používané vo financiách a ekonometrii. Jej význam potvrdzuje nielen široké uplatnenie v klasických ekonomických modeloch ako je CAPM [55] či jeho päťfaktorové rozšírenie [20], ale aj jej využitie v oblasti relatívneho oceňovania [30].

Využívanie SVM vo financiách má svoju dlhoročnú tradíciu – už na prelome tisícročí vznikali články, ktoré demonštrovali ich aplikáciu pri riešení rôznych finančných problémov ([56], [7]). Hoci využitie SVR v kontexte relatívneho oceňovania nie je v súčasnej literatúre výrazne preskúmané, rozhodli sme sa tento model zahrnúť vzhľadom na jeho širokú rozšírenosť a vysokú flexibilitu.

Posledným modelom je XGB, ktorý je spomedzi modelov najnovší. Algoritmus gradient boostingu bol vyvinutý Friedmanom pred viac ako štvrťstoročím [23]. Zaujímavosťou ale je, že XGB samotný vznikol až v roku 2014 a sprvu bol ignorovaný laickou verejnosťou, ale i akademickou obcou [44]. Od roku 2016, kedy sa dostal do povedomia, je využívaný v rôznych oblastiach, o čom sa možno presvedčiť napr. v [46] a [18].

Vhodnosť modelu XGB pre ekonomické a finančné aplikácie potvrdzujú aj viaceré akademické články ([25], [51]), ktoré dokumentujú jeho úspešné využitie pri relatívnom oceňovaní firiem a odhadovaní poklesu či nárastu ceny akcií.

Výberom týchto troch metód pokrývame spektrum od základného benchmarku, cez flexibilnú nelineárnu techniku až po moderný a výpočtovo efektívny algoritmus.

2.1 Lineárny regresný model

Lineárna regresia (LR) predstavuje základný štatistický nástroj na skúmanie závislosti medzi závislou a nezávislými premennými. K výhodám využitia lineárnej regresie patrí

jednoduchosť implementácie a priamočiara interpretácia získaných výsledkov. Získané výsledky využijeme ako benchmark pre porovnanie s ostatnými modelmi.

V tejto časti, vychádzajúc z [34], stručne popíšeme matematický model lineárnej regresie a metodiku odhadu parametrov.

2.1.1 Špecifikácia modelu

Majme dátovú sadu

$$\mathcal{D} = \{x_i, y_i\}(|D| = n, x_i \in \mathbb{R}^p, y_i \in \mathbb{R}). \tag{36}$$

Lineárny regresný model opisuje vzťah medzi závislou premennou y_i a vektorom prediktorov $x_i \in \mathbb{R}^p$

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \sim \text{i.i.d. } \mathcal{N}(0, \sigma^2), \tag{37}$$

kde $\beta = (\beta_0, \ldots, \beta_p)^T$ predstavuje vektor parametrov a ε_i je chybový člen. V maticovom tvare dostávame

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n). \tag{38}$$

Základné predpoklady modelu zahŕňajú linearitu, nezávislosť, homoskedasticitu a normalitu reziduí. Aj bez splnenia všetkých týchto podmienok mnohé vlastnosti zostávajú zachované [34].

2.1.2 Odhad parametrov

Parametre β odhadujeme minimalizáciou kvadratickej chybovej funkcie

$$Q(\beta) = (y - X\beta)^T (y - X\beta). \tag{39}$$

Riešením rovnice $X^T X\beta = X^T y$ dostaneme odhad:

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{40}$$

Predikované hodnoty potom zodpovedajú $\hat{y} = X\hat{\beta}$.

Lineárna regresia má elegantnú geometrickú interpretáciu ako ortogonálnu projekciu vektora y na podpriestor $\mathcal{C}(X)$, ktorý je generovaný stĺpcami matice X. Odhady $\hat{y}$ sú ortogonálne na reziduá $r = y - \hat{y}$.

Odhad $\hat{\beta}$ je nielen nevychýlený, ale podľa Gauss-Markovovej vety aj najlepší lineárny nevychýlený odhad (BLUE). Platí

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}, \tag{41}$$

$$\hat{\sigma}^2 = \frac{\|y - X\hat{\beta}\|^2}{n - p - 1}. \tag{42}$$

Pri normálnej chybe ε má odhad $\hat{\beta}$ normálne rozdelenie, umožňujúce budovanie intervalov spoľahlivosti a testovanie štatistických hypotéz.

2.2 Teória Support Vector Machines

Druhým modelom v poradí je algoritmus založený na *Support Vector Machines* (SVM). Hoci SVM patria medzi najvýkonnejšie klasifikačné algoritmy, ich základnú štruktúru možno vhodnou úpravou prispôsobiť na regresné účely. Takto modifikovaný model označujeme ako *Support Vector Regression* (SVR). Tento prístup umožňuje zachytiť nelineárne vzťahy medzi nezávislými premennými a závislou premennou pomocou jadrových funkcií. Táto časť vychádza z myšlienok, ktoré sú prezentované v článkoch [6] a [24]. Hlavným princípom SVM modelu je hľadanie optimálnej hranice (tzv. nadroviny), ktorá dokáže najlepšie oddeliť dve alebo viac tried dátových bodov v určitom priestore. Zásadnou výhodou SVM je, že sa pri hľadaní tejto hranice sústreďuje iba na tzv. *support* vektory, teda tie dáta, ktoré sa nachádzajú najbližšie k deliacim plochám. Týmto prístupom sa eliminuje vplyv nepotrebných bodov, čo zvyšuje robustnosť modelu.

Opäť majme dátovú sadu v zmysle (36). Matematicky je problém definovaný ako snaha nájsť nadrovinu v priestore tak, aby vzdialenosť od tejto nadroviny k najbližším bodom oboch tried (tzv. *margins*) bola maximálna. Formálne to možno zapísať ako optimalizačný problém

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{pri podmienkach} \quad y_i(w^T x_i + b) \geq 1, \ \forall i, \tag{43}$$

kde w predstavuje vektor normály nadroviny, b je posunutie nadroviny a $y_i \in \{-1, 1\}$ označuje príslušnosť i-teho bodu k jednej z dvoch tried. Výsledná klasifikačná funkcia potom nadobúda tvar

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i x_i^T x + b\right), \tag{44}$$

kde α_i sú Lagrangeove multiplikátory, ktoré sú nenulové práve pre support vektory, x je vektor prediktorov pozorovania, ktoré chceme klasifikovať.

V prípade, že sú dáta nelineárne separovateľné, pôvodná SVM metóda umožňuje ich spracovanie pomocou zavedenia tzv. slack premenných ξ_i, ktoré umožňujú drobné chyby v klasifikácii. Takto vzniká tzv. *soft-margin* SVM

$$\min_{w,\xi,b} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \right\}, \quad \text{za podmienok} \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0. \tag{45}$$

Parameter C pritom kontroluje, ako veľmi model penalizuje chyby.

Významnou vlastnosťou SVM je možnosť ich aplikácie na nelineárne separovateľné dáta bez explicitnej transformácie do vyšších dimenzií. Toto sa realizuje prostredníctvom kernelových metód, kde skalárne súčiny v novom, viacdimenzionálnom priestore možno vyjadriť pomocou kernelovej funkcie $K(x_i, x)$

$$K(x_i, x) = \langle \phi(x_i), \phi(x) \rangle, \tag{46}$$

kde $\phi(x)$ predstavuje transformáciu pôvodných dát do nového priestoru. V praxi sa najčastejšie používajú napríklad lineárny, polynomiálny alebo Gaussov (RBF) kernel. Rozhodovacia funkcia v kernelovej forme má potom tvar

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \right). \tag{47}$$

Kým klasické SVM riešia klasifikačné úlohy, ich princíp možno analogicky využiť aj pri regresných problémoch, čím vzniká metóda známa ako *Support Vector Regression* (SVR) [1]. Cieľom SVR je nájsť funkciu $f(x)$, ktorá dobre aproximuje spojitú cieľovú premennú a zároveň je čo najhladšia. Hlavný rozdiel oproti klasifikácii spočíva v tom, že SVR namiesto hľadania maximálneho marginu medzi triedami hľadá takú funkciu, ktorá sa čo najmenej odchýli od skutočných hodnôt cieľovej premennej v rámci preddefinovanej tolerancie $\varepsilon > 0$. Regresný problém je teda formulovaný ako

$$\min_{w,\xi_i,\xi_i^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*), \tag{48}$$

s podmienkami

$$y_i - (w^T\phi(x_i) + b) \leq \varepsilon + \xi_i, \tag{49}$$

$$(w^T\phi(x_i) + b) - y_i \leq \varepsilon + \xi_i^*, \tag{50}$$

$$\xi_i, \xi_i^* \geq 0. \tag{51}$$

Výsledná regresná funkcia má tvar

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b, \tag{52}$$

kde znova α_i a α_i^* sú Lagrangeove multiplikátory.

2.3 Extreme Gradient Boosting (XGB)

Tretím modelom je algoritmus extrémneho gradientného boostingu. Detailnejšie a komplexnejšie spracovanie teoretických základov a matematických postupov algoritmu XGB možno nájsť v zdroji [33]. V rámci tejto práce sme niektoré technické kroky a podrobnosti vynechali s cieľom zachovať prehľadnosť výkladu, no uvedené vzťahy a argumentácia priamo vychádzajú práve zo spomínaného článku. XGB, založený na princípe gradient boosting rozhodovacích stromov, patrí medzi najefektívnejšie algoritmy strojového učenia. Bol vyvinutý s cieľom maximalizovať predikčnú presnosť a minimalizovať výpočtovú náročnosť pri analýze veľkých a komplexných dátových súborov.

Majme dátovú sadu v zmysle (36). Predikcia modelu $\hat{y}_i$ je definovaná ako súčet predikcií K stromov f_k, čo sa dá vyjadriť ako

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}, \tag{53}$$

kde $\mathcal{F}$ reprezentuje množinu všetkých stromov f_k, ktorú zapisujeme

$$\mathcal{F} = \left\{ f(x) = w_{q(x)} \,\middle|\, q : \mathbb{R}^p \to T,\ w \in \mathbb{R}^T \right\}. \tag{54}$$

Strom f_k priradzuje vektoru prediktorov $x \in \mathbb{R}^p$ hodnotu $w_q(x)$ prislúchajúca konkrétnemu listu zo stromu. Listy (angl. *leaves*) predstavujú koncové uzly (*splity*) stromu, ktoré neobsahujú ďalšie rozdelenia a v ktorých sa nachádzajú finálne predikcie stromu, T označuje počet listov v strome.

XGB optimalizuje nielen presnosť predikcie, ale aktívne reguluje aj zložitosť stromov. Minimalizačná funkcia je definovaná ako

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{55}$$

kde $l(y_i, \hat{y}_i)$ je stratová funkcia a $\Omega(f_k)$ predstavuje penalizáciu komplexnosti jednotlivých stromov. Táto regularizačná funkcia má tvar

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2, \tag{56}$$

kde $\gamma > 0$ kontroluje komplexnosť stromu a $\lambda > 0$ penalizuje veľkosť predikcií v listoch. Podobná regularizačná technika bola použitá aj v modeli *Regularized Greedy Forest (RGF)*. XGB je však ľahšie paralelizovateľný a minimalizovaná funkcia je jednoduchšia [33].

Množstvo parametrov v modeli spôsobuje, že minimalizovaná funkcia nepatrí medzi triedy funkcií, ktorých minimum sa dá efektívne nájsť metódami nelineárneho programovania. Namiesto toho sa XGB trénuje aditívne. Nech $\hat{y}_i^{(t)}$ označuje predikciu pre i-te pozorovanie v t-tej iterácii. Aby sme model ďalej zlepšili (t.j. zmenšili jeho stratovú funkciu), pridávame v každom kroku novú funkciu f_t, ktorá minimalizuje funkciu

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t). \tag{57}$$

Týmto spôsobom do modelu chamtivo (po anglicky sa používa slovo *greedily*) pridávame takú funkciu f_t, ktorá najviac zlepší výkonnosť modelu podľa (55). Na efektívnu optimalizáciu cieľovej funkcie možno v praxi použiť druhý rád Taylorovho rozvoja stratovej funkcie, čo urýchľuje výpočet. Dostávame

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t), \tag{58}$$

kde $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ a $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ sú prvé a druhé derivácie stratovej funkcie vzhľadom na predikcie z predchádzajúcej iterácie. Tieto hodnoty predstavujú gradienty, ktoré sa využívajú pri výstavbe ďalšieho stromu. Konštantné členy v optimalizačnom výraze možno vynechať, čím získame zjednodušený tvar funkcie pre iteráciu t. Pri vytváraní nového stromu v každej iterácii teda dostávame zjednodušenú funkciu

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t), \tag{59}$$

kde g_i predstavuje gradient a h_i Hessovu maticu stratovej funkcie vzhľadom na súčasnú predikciu.

Optimalizácia stromu sa vykonáva pomocou rozdelenia dátových bodov do listov $I_j = \{i | q(x_i) = j\}$ s predikciami w_j, pričom optimalizovanou cieľovou funkciou sa stáva

$$\tilde{\mathcal{L}}^{(t)} = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T. \tag{60}$$

Optimálna predikcia v liste je určená analyticky vzťahom

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \tag{61}$$

a zodpovedajúca cieľová funkcia po vytvorení stromu je

$$\tilde{\mathcal{L}}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T. \tag{62}$$

Výsledkom iteratívneho algoritmu je K stromov, pomocou ktorých možno v zmysle (53) urobiť predikcie.

3 Automatizovaný model

V tejto časti práce sa zameriavame na praktickú implementáciu automatizovaného prístupu k oceňovaniu firiem. Cieľom je vytvoriť automatizovaný model založený na princípoch relatívneho oceňovania, pričom využijeme tri regresné modely predstavené v predchádzajúcej časti. Tieto modely budú slúžiť na predikciu cieľových premenných $m2b$, $ev2a$ a $ev2s$, nasledujúc myšlienky v [25]. Zostáva zodpovedať poslednú otázku: aký typ dát zvoliť ako prediktory. V súlade s moderným prístupom v literatúre, kde sa často využívajú účtovné ukazovatele ([28], [30], [25]), sme sa rozhodli ísť rovnakou cestou. V nasledujúcich častiach bližšie predstavíme použitý dátový súbor a následne kvantitatívne vyhodnotíme výkonnosť jednotlivých regresných modelov.

Poznámka 3.0.1. Podobne ako autori v [25] z dôvodu prehľadnosti nebudeme používať časový index t. V prípade, že budeme pracovať s rôznymi veličinami, vždy budeme predpokladať, že sa vzťahujú na to isté časové obdobie.

Ako sme uviedli v časti venovanej relatívnemu oceňovaniu, prvým krokom by mal byť výber vhodnej skupiny porovnateľných firiem. Autori v [25] však tento krok vynechávajú s odôvodnením, že ak pri oceňovaní využívame metódy strojového učenia založené na účtovných ukazovateľoch, samotný algoritmus by mal vybrať relevantnú skupinu firiem na základe podobnosti v dátach. Takýto prístup nielen zjednodušuje celý proces, ale zároveň napomáha eliminovať subjektívne zásahy analytika či investora — čo je aj jedným z cieľov tejto práce. Preto tento postup volíme aj my.

3.1 Špecifikácia dát

Dáta na regresiu sme získali pomocou API rozhrania služby Financial Modeling Prep [22]. Stiahli sme účtovné výkazy (súvahu, výkaz ziskov a strát, výkaz peňažných tokov) spoločností, ktoré boli k dátumu 31.12.2024 v indexe S&P 500 na ročnej báze. Analyzovaný dátový súbor pokrýva obdobie od 1. januára 1985 do 24. decembra 2024.

Poznámka 3.1.1. Našim zámerom bolo spracovať štvrťročné výkazy s cieľom maximalizovať veľkosť dátovej vzorky. Tento cieľ sa však nepodarilo naplniť, keďže niektoré dáta boli reportované na štvrťročnej báze, iné na štvrťročnej kumulovanej báze. Z dôvodu, že

nebolo možné skontrolovať všetky záznamy manuálne, rozhodli sme sa pracovať s dátami na ročnej báze.

Poznámka 3.1.2. Prirodzene, dáta neboli dostupné pre všetky spoločnosti počas celého sledovaného obdobia. Niektoré firmy v danom čase ešte neboli súčasťou indexu, iné dokonca v časti obdobia vôbec neexistovali.

Pred samotným zostavením modelu bolo nevyhnutné zabezpečiť úplnosť údajov, keďže niektoré kľúčové veličiny neboli v databáze dostupné v dostatočnom rozsahu. To bol aj prípad trhovej kapitalizácie (MC), ktorá bola potrebná na výpočet $m2b$, no poskytovateľ dát ju sprístupňoval iba do roku 2020. Z tohto dôvodu sme pristúpili k jej rekonštrukcii podľa nasledujúceho postupu

$$m2b = \frac{mEV}{BV} = \frac{close\ price \times \frac{NI}{\text{čistý zisk na akciu}}}{BV}, \tag{63}$$

kde *close price* je cena akcie na konci obchodovacieho dňa a čistý zisk na akciu rátame ako $\frac{NI}{počet\ emitovaných\ akcií}$. Prehľad prediktorov aj s ich výpočtom uvádzame v Tab. 1.

Prediktor	**Výpočet**
ROE	$\frac{NI}{BV}$
ROA	$\frac{NI}{A}$
Operačná marža	$\frac{OPINC}{S}$
Zisková marža	$\frac{NI}{S}$
CF marža	$\frac{\text{FCFF}}{S}$
Pomer dlhu a vlastného imania	$\frac{Debt}{BV}$
ATR	$\frac{S}{A}$
EBITDA marža	$\frac{\text{EBITDA}}{S}$
day_count	$\frac{date}{\text{počet dní analyzovaného obdobia}}$

Tab. 1: Zoznam prediktorov a ich výpočet.

Symbol $OPINC$ v Tab. 1 reprezentuje prevádzkový zisk. Pri výbere prediktorov sme vychádzali z metodiky uvedenej v [25], avšak kvôli rozdielnej štruktúre dát sme nedokázali dosiahnuť rovnaký počet prediktorov. Vybrali sme preto iba tie, k výpočtu ktorých sme mali dostupné údaje. Novátorskou myšlienkou, ktorá sa metodicky odkláňa od článku [25]

je zadefinovanie časovej zložky *day_count* ako prediktora. Pre tento výber sme sa rozhodli na základe myšlienky prezentovanej v [12]: relatívne násobky v čase rastú, zachytenie časovej zložky môže byť preto zaujímavé. Prediktor *day_count*, ktorý reprezentuje relatívne časové poradie pozorovania v rámci analyzovaného obdobia, bol získaný z poradového čísla dňa od začiatku dátového súboru (*date*) normovaného na interval $[0, 1]$.

3.1.1 Obmedzenia dát

Najväčším problémom spojeným s dátovou sadou bola nekonzistentnosť údajov. Často sa stávalo, že v jednom zázname boli náklady reportované so záporným znamienkom, inokedy s kladným. Táto chyba sa potom prenášala na všetky výpočty, čo neumožňovalo efektívnu a správnu analýzu. Veľmi častým javom bolo aj to, že sa v súvahe nerovnali aktíva a pasíva. Zjavne chybné údaje sme automaticky vyradili. Po úpravách nám zostalo 8317 pozorovaní, čo je necelých 53% pôvodných dát.

3.1.2 Odstraňovanie odľahlých pozorovaní a príprava dát

Na vylepšenie výsledkov regresie bolo potrebné odstrániť odľahlé pozorovania. Použili sme dva prístupy. Prvým bolo odstránenie odľahlých pozorovaní pomocou medzikvartilového rozpätia (IQR) [48]. Pri tejto technike sa na každý stĺpec dát pozeráme samostatne a odstraňujeme všetky pozorovania, ktoré obsahujú aspoň jednu hodnotu ležiacu mimo rozsahu

$$\mathrm{Q1} - 1{,}5 \times \mathrm{IQR} \quad \text{až} \quad \mathrm{Q3} + 1{,}5 \times \mathrm{IQR},$$

kde Q1 predstavuje prvý kvartil a Q3 predstavuje tretí kvartil daného stĺpca. Hoci ide o štandardný spôsob identifikácie extrémnych hodnôt, v praxi tento prístup vedie k výraznej redukcii počtu pozorovaní.

Druhý prístup je inšpirovaný analýzou hlavných komponentov (PCA), ktorá sa štandardne využíva na redukciu dimenzie dát [38] [35]. V našom prípade však jej princípy aplikujeme na účely identifikácie odľahlých pozorovaní vo viacrozmernom priestore. Konkrétne, dáta preškálujeme a vycentrujeme. Následne každému pozorovaniu priradíme euklidovskú vzdialenosť od stredu v PCA priestore. Na základe rozdelenia týchto vzdialeností určujeme hranicu odľahlosti ako 90. percentil a pozorovania, ktoré túto hranicu prekročia, z analýzy odstránime.

Poznámka 3.1.3. Prediktor *day_count* prirodzene do odstraňovania odľahlých pozorovaní nezahŕňame.

Oba prístupy generujú rozdielne množiny, je teda cenné si prístupy a výsledky, ktoré sú s nimi spojené, porovnať. Ďalej budeme o týchto dvoch prístupoch hovoriť ako o filtráciách. Dáta sme ešte rozdelili na tri časti: trénovacie (od roku 1985 do roku 2022), testovacie (2022-2023) a dáta, ktoré používame na konštrukciu portfólia (2023-2024). Testovacie dáta sme použili na vyhodnotenie výkonnosti modelov. Vzhľadom na to, že všetky cieľové premenné boli zošikmené na pravú stranu, aplikovali sme na ne logaritmickú transformáciu. Logaritmicky transformované cieľové premenné budeme značiť ln $m2b$, ln $ev2a$ a ln $ev2s$.

Poznámka 3.1.4. Len pre zaujímavosť, IQR skutočne vyfiltrovalo obrovské množstvo dát. Po IQR filtrácii sa dostávame na číslo 5106. Keďže pri PCA odstraňujeme iba 10% dát, vo finále sme dostali 7223 záznamov. Niektoré pozorovania sme odstránili samotnou logaritmickou transformáciou, ak boli cieľové premenné záporné.

3.2 Numerické výsledky

V tejto časti sa venujeme výsledkom regresie všetkých troch modelov na všetkých cieľových premenných. Metriky, ktorými meriame presnosť sú nasledovné:

Definícia 3.2.1. Nech y_i je skutočná hodnota cieľovej premennej a $\hat{y}_i$ je odhad, $i = 1, \ldots, n$ označuje index pozorovania. Sumy štvorcov chýb (SSE) definujeme nasledovne

$$SSE = \sum_{i=1}^{n} \Big(y_i - \hat{y}_i\Big)^2. \tag{64}$$

Definícia 3.2.2. Nech platia označenia z definície 3.2.1. Celkovú sumu štvorcov (SST) definujeme nasledovne

$$SST = \sum_{i=1}^{n} \Big(y_i - \bar{y}\Big)^2, \tag{65}$$

kde $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Definícia 3.2.3. Nech platia predpoklady z 3.2.1 a 3.2.2. Koeficient determinácie (R^2) definujeme nasledovne

$$R^2 = 1 - \frac{SSE}{SST}. \tag{66}$$

Definícia 3.2.4. Nech platia predpoklady z 3.2.1. Absolútnu mediánovú chybu $MedAE$ definujeme ako

$$MedAE = \text{median}\left(\{\,|y_i - \hat{y}_i|\,\}_{i=1}^{n}\right), \tag{67}$$

a relatívnu mediánovú chybu $MedRE$ definujeme ako

$$MedRE = \text{median}\left(\{\frac{|y_i - \hat{y}_i|}{|y_i|}\}_{i=1}^{n}\right). \tag{68}$$

Definícia 3.2.5. Nech platia predpoklady z 3.2.2. Odmocninu zo strednej kvadratickej chyby $RMSE$ definujeme ako

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(y_i - \hat{y}_i\big)^2}. \tag{69}$$

Metriky (66), (67), (68) a (69) predstavujú štandardný nástroj na meranie výkonnosti regresných modelov [25].

Cieľová premenná	**Model**	R^2	$MedAE$	$MedRE$	$RMSE$
ln $m2b$	LR	0,091	1,044	0,368	2,381
ln $m2b$	XGB	0,157	0,935	0,345	2,293
ln $m2b$	SVR	0,180	1,106	0,318	2,262
ln $ev2a$	LR	0,258	0,432	0,305	0,960
ln $ev2a$	XGB	0,358	0,404	0,294	0,892
ln $ev2a$	SVR	0,344	0,377	0,273	0,902
ln $ev2s$	LR	0,274	0,793	0,299	2,152
ln $ev2s$	XGB	0,363	0,635	0,254	2,016
ln $ev2s$	SVR	0,397	0,800	0,263	1,961

Tab. 2: Regresné metriky pre IQR filtrovanie.

Ako vidíme v Tab. 2, lineárna regresia dosahuje v porovnaní s nelineárnymi modelmi (XGB a SVR) systematicky nižšie hodnoty koeficientu determinácie R^2, čo naznačuje jej obmedzenú schopnosť zachytiť zložitejšie vzťahy medzi prediktormi a cieľovými premennými. Pri premennej ln $ev2s$ sú všetky modely relatívne nepresné, keďže najvyššie R^2 dosiahlo len 0,180. Napriek tomu XGB aj SVR vykazujú nižšie hodnoty chýb $MedAE$,

$MedRE$ a $RMSE$ než lineárna regresia, čo svedčí o lepšom výkone a menších odchýlkach. Premenná ln $ev2a$ je podstatne lepšie predikovateľná – najvyššie $R^2 = 0{,}358$ dosahuje XGB, ktorý má zároveň najnižšie $RMSE = 0{,}892$, zatiaľ čo SVR vykazuje najnižšie hodnoty $MedAE$ aj $MedRE$. Podobná situácia nastáva pri premennej ln $ev2s$, kde je najúspešnejším modelom SVR s najvyšším $R^2 = 0{,}397$ a najnižším $RMSE = 1{,}961$, pričom XGB vykazuje najnižšiu relatívnu chybu ($MedRE = 0{,}254$). Z výsledkov možno usúdiť, že nelineárne modely podávajú stabilne lepší výkon a lepšie zachytávajú vzťahy v dátach po aplikácii IQR filtrovania.

Cieľová premenná	**Model**	R^2	$MedAE$	$MedRE$	$RMSE$
ln $m2b$	LR	0,281	1,312	0,401	2,675
ln $m2b$	XGB	0,539	1,022	0,311	2,142
ln $m2b$	SVR	0,512	0,989	0,313	2,205
ln $ev2a$	LR	0,380	0,438	0,325	0,953
ln $ev2a$	XGB	0,542	0,325	0,250	0,819
ln $ev2a$	SVR	0,582	0,346	0,232	0,783
ln $ev2s$	LR	0,438	0,957	0,309	2,273
ln $ev2s$	XGB	0,676	0,714	0,240	1,727
ln $ev2s$	SVR	0,576	0,898	0,288	1,973

Tab. 3: Regresné metriky pre PCA filtrovanie.

Po aplikácii PCA filtrovania sa výrazne zlepšila predikčná schopnosť všetkých modelov v porovnaní s IQR prístupom, o čom sa možno presvedčiť v Tab. 3. Pri premennej ln $ev2s$ dosahuje najvyššie $R^2 = 0{,}539$ model XGB, nasledovaný SVR ($R^2 = 0{,}512$), pričom oba nelineárne modely výrazne prekonávajú lineárnu regresiu ($R^2 = 0{,}281$). Najnižšiu absolútnu chybu má SVR ($MedAE = 0{,}989$), zatiaľ čo najnižšiu relatívnu chybu dosahuje XGB ($MedRE = 0{,}311$). Premenná ln $ev2a$ je predikovaná ešte lepšie – SVR dosahuje najvyššie $R^2 = 0{,}582$ a najnižšie hodnoty $RMSE = 0{,}783$ aj $MedRE = 0{,}232$. XGB má síce mierne nižšie $R^2 = 0{,}542$, ale najnižšie $MedAE = 0{,}325$. Lineárna regresia v tomto prípade výrazne zaostáva vo všetkých metrikách. Pri premennej ln $ev2s$ dosahuje najvyššie $R^2 = 0{,}676$ opäť XGB, ktorý zároveň dominuje v chybových metrikách: $RMSE = 1{,}727$, $MedAE = 0{,}714$ a $MedRE = 0{,}240$. Hoci SVR zaostáva v R^2 aj $RMSE$, stále si udr-

žuje relatívne priaznivé hodnoty chýb. Výsledky jednoznačne ukazujú, že PCA filtrácia zvyšuje predikčnú presnosť modelov, pričom nelineárne prístupy, najmä XGB, sú výrazne presnejšie.

3.2.1 Intepretácia výsledkov

Vzhľadom na nízku hodnotu R^2 pre lineárnu regresiu uvádzame regresné koeficienty súhrnne v Tab. 4 a Tab. 5 a nevenujeme im väčšiu pozornosť.

Premenná	**ln m2b**		**ln ev2a**		**ln ev2s**	
	Koef. β_i	p-value	Koef. β_i	p-value	Koef. β_i	p-value
Intercept	1,15	0	0,77	0	0,99	0
ROE	1,24	0	−1,33	0	−1,72	0
ROA	3,66	0	7,29	0	4,48	0
Operačná marža	0,58	0	0,24	0,04	0,18	0,20
Zisková marža	−1,96	0	−1,32	0	0,54	0,007
CF marža	1,19	0	0,71	0	1,06	0
Pomer dlhu a vlastného imania	0,00	0,83	0,06	0	0,09	0
ATR	0,14	0	0,10	0	−0,26	0
EBITDA marža	−0,37	0,004	0,12	0,18	0,59	0
day_count$_t$	0,76	0	0,43	0	0,59	0

Tab. 4: Koeficienty lin. regresie pre IQR filtrovanie pre cieľové premenné ln $m2b$, ln $ev2a$ a ln $ev2s$.

Jednou z kľúčových výhod lineárnej regresie, ktorú sme v teoretickej časti zdôraznili, je jej vysoká miera interpretovateľnosti. Naopak, pokročilé algoritmy strojového učenia sú často označované ako tzv. *black-box* modely: čierne skrinky, ktoré síce dosahujú vysokú predikčnú presnosť, no spôsob, akým dochádzajú k svojim rozhodnutiam, zostáva pre používateľa netransparentný [54]. Pri aplikáciách, ako je oceňovanie firiem, taktiež nestačí poznať len samotnú predikciu. Rovnako dôležité je pochopiť, na základe ktorých vstupov model dospel k svojmu záveru.

Aj keď je XGB považovaný za komplexný model, je postavený na rozhodovacích stromoch, čo umožňuje aspoň čiastočnú interpretáciu jeho výstupov. Model eviduje, ktoré pre-

Premenná	**ln m2b**		**ln ev2a**		**ln ev2s**	
	Koef. β_i	p-value	Koef. β_i	p-value	Koef. β_i	p-value
Intercept	1,22	0	0,80	0	1,10	0
ROE	0,71	0	−0,40	0	−0,50	0
ROA	3,41	0	4,97	0	2,33	0
Operačná marža$_t$	0,15	0,053	−0,02	0,75	−0,10	0,12
Zisková marža$_t$	−1,15	0	−0,90	0	0,08	0,30
CF marža$_t$	0,41	0	0,31	0	0,55	0
Pomer dlhu a vlastného imania$_t$	0,09	0	0,02	0	0,03	0
ATR_t	0,15	0	0,05	0	−0,37	0
EBITDA marža$_t$	0,04	0,41	0,21	0	0,86	0
day_count$_t$	0,76	0	0,38	0	0,59	0

Tab. 5: Koeficienty lin. regresie po PCA filtrovaní pre cieľové premenné ln $m2b$, ln $ev2a$ a ln $ev2s$.

menné boli použité pri rozdeľovaní uzlov (tzv. *splits*) a aký vplyv mali tieto rozhodnutia na celkovú predikčnú výkonnosť. Na tento účel poskytuje tri základné metriky dôležitosti: zlepšenie (*gain*), frekvencia (*frequency*) a pokrytie (*cover*).

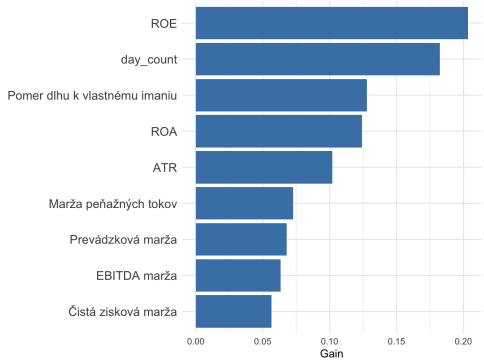
Zlepšenie vyjadruje relatívny príspevok prediktora k celkovému zníženiu chyby modelu. Predstavuje podiel súčtu všetkých redukcií chyby v uzloch, kde sa daný prediktor použil, na súčte redukcií všetkých prediktorov. Zachytáva, do akej miery prediktor zvyšuje presnosť modelu: aj pri zriedkavom výskyte môže mať vysoké zlepšenie, ak významne zníži chybu. Považuje sa preto za najinformatívnejšiu metriku pri hodnotení významnosti.

Frekvencia udáva, ako často sa prediktor použil pri rozdeľovaní uzlov vo všetkých stromoch, relatívne voči ostatným. Označuje teda, ktoré prediktory sa v modeli vyskytujú najčastejšie, no nezohľadňuje ich vplyv na presnosť. Prediktor s vysokou frekvenciou tak nemusí výrazne prispievať k zníženiu chyby. Táto metrika je preto vhodná najmä na analýzu štruktúry modelu, nie jeho výkonnosti.

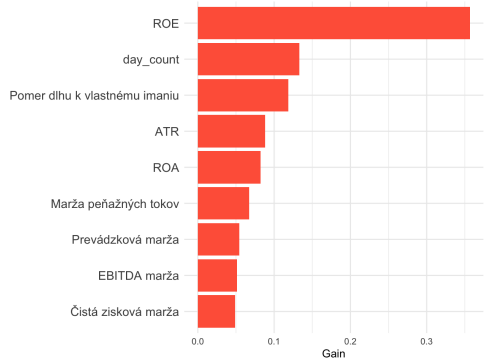
Pokrytie udáva relatívny priemerný počet pozorovaní, ktoré prejdú uzlami, kde sa daný prediktor použil na rozdelenia uzla. Vyjadruje teda rozsah, do akej miery prediktor zasiahol dátovú množinu. Vysoké pokrytie zvyčajne dosahujú premenné používané v

skorších rozhodovacích uzloch, no metrika nehodnotí ich vplyv na presnosť, iba rozsah ich pôsobenia. Pri interpretácii výstupov z XGB je preto vhodné kombinovať tieto tri metriky, pričom ako hlavné kritérium sa najčastejšie používa zlepšenie.

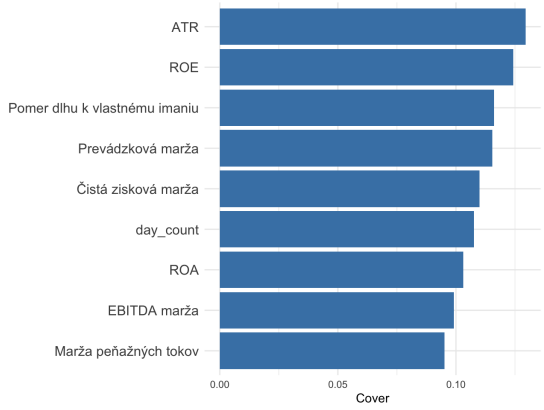
Pri analýze regresie cieľovej premennej ln $m2b$, vychádzajúcej z Obr. 1, Obr. 2, Obr. 3, Obr. 4, Obr. 5 a Obr. 6, možno konštatovať, že najvyššie zlepšenie pri IQR dosahuje ROE, za ňou nasleduje prediktor day_count a pomer dlhu k vlastnému imaniu. Po aplikácii PCA filtrovania zostáva poradie v zlepšení nezmenené, no v pokrytí sa na prvé miesto dostáva ATR a day_count sa posúva na tretiu priečku. Pre frekvenciu zostávajú ROE a pomer dlhu k vlastnému imaniu najčastejšie využívané. Maržové dosahujú nízke hodnoty vo všetkých troch metrikách, čo poukazuje na ich obmedzený význam.



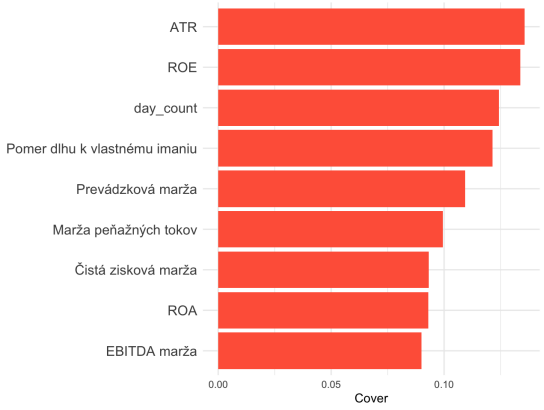
Obr. 1: Zlepšenie pre XGB podľa prediktorov, IQR filtrácia, premenná ln $m2b$.



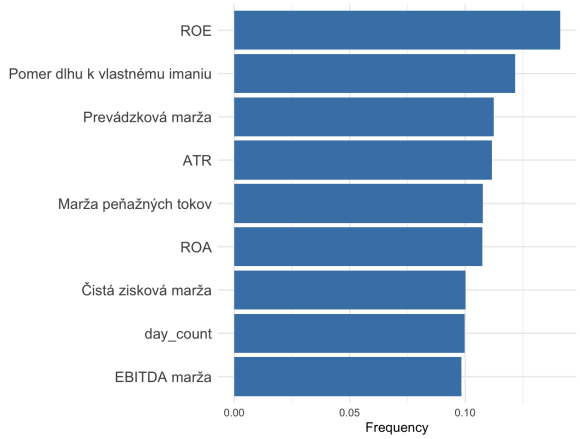
Obr. 2: Zlepšenie pre XGB podľa prediktorov, PCA filtrácia, premenná ln $m2b$.



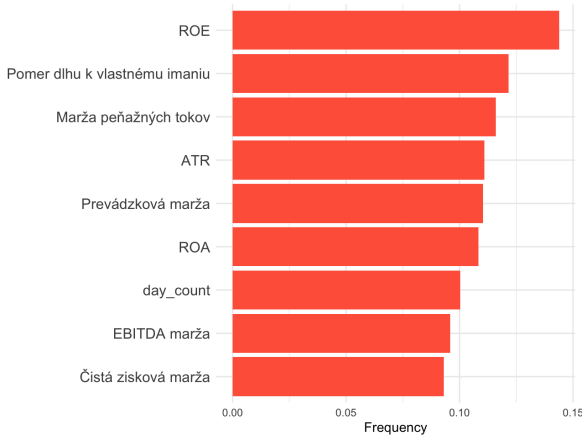
Obr. 3: Pokrytie prediktorov pre XGB, IQR filtrácia, premenná ln $m2b$.



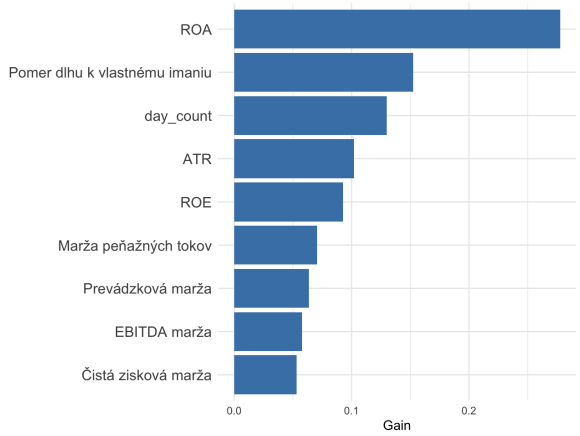
Obr. 4: Pokrytie prediktorov pre XGB, PCA filtrácia, premenná ln $m2b$.



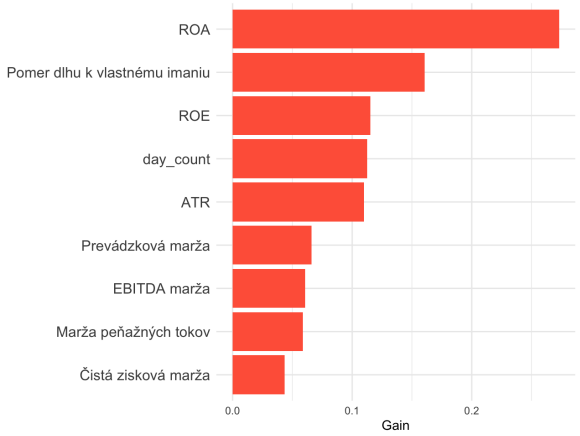
Obr. 5: Frekvencia prediktorov pre XGB, IQR filtrácia, premenná ln $m2b$.



Obr. 6: Frekvencia prediktorov pre XGB, PCA filtrácia, premenná ln $m2b$.

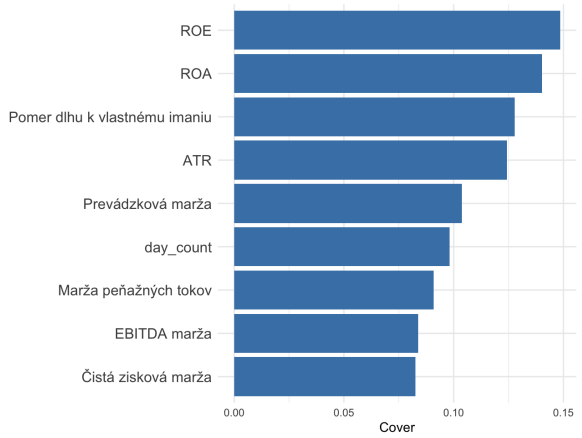


Obr. 7: Zlepšenie pre XGB podľa prediktorov, IQR filtrácia, premenná ln $ev2a$.

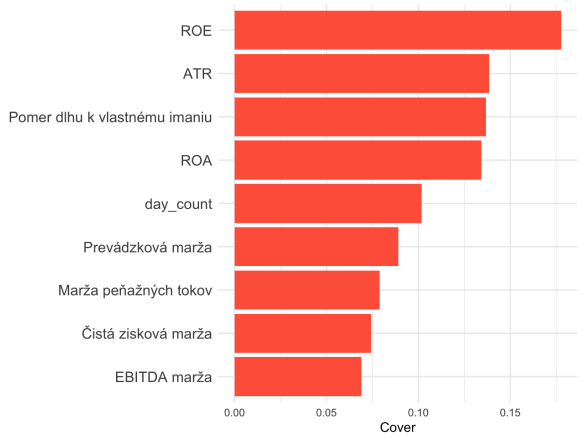


Obr. 8: Zlepšenie pre XGB podľa prediktorov, PCA filtrácia, premenná ln $ev2a$.

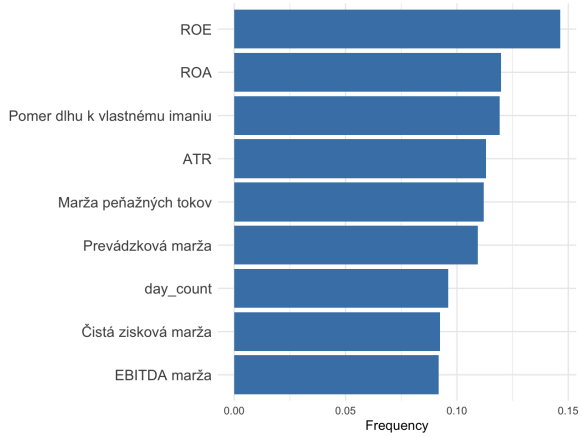
Na základe Obr. 7, Obr. 8, Obr. 9, Obr. 10, Obr. 11 a Obr. 12 možno konštatovať, že pre cieľovú premennú ln $ev2a$ vykazujú najvyššie hodnoty zlepšenia prediktory ROA a pomer dlhu k vlastnému imaniu. Pri filtrovaní pomocou PCA sa v metrikách pokrytia a frekvencie dodatočne zviditeľňuje aj ukazovateľ ATR. Naprieč oboma typmi filtrovania sa ROE konzistentne nachádza medzi najčastejšie využívanými prediktormi podľa frekvencie aj pokrytia. Maržové ukazovatele sa naopak v rámci všetkých troch metrík javia ako menej významné. Medzi filtráciami IQR a PCA nepozorujeme výrazné rozdiely v identifikovaných relevantných prediktoroch.



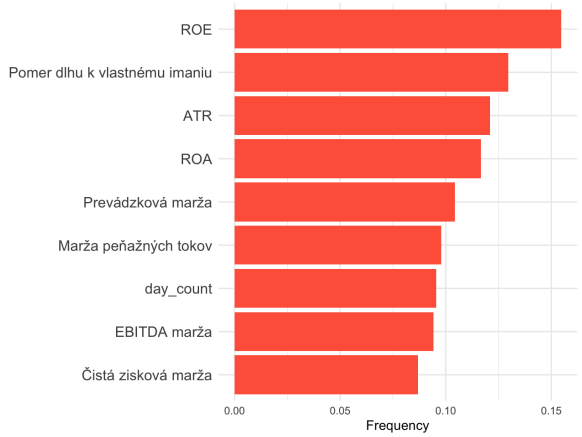
Obr. 9: Pokrytie prediktorov pre XGB, IQR filtrácia, premenná ln $ev2a$.



Obr. 10: Pokrytie prediktorov pre XGB, PCA filtrácia, premenná ln $ev2a$.



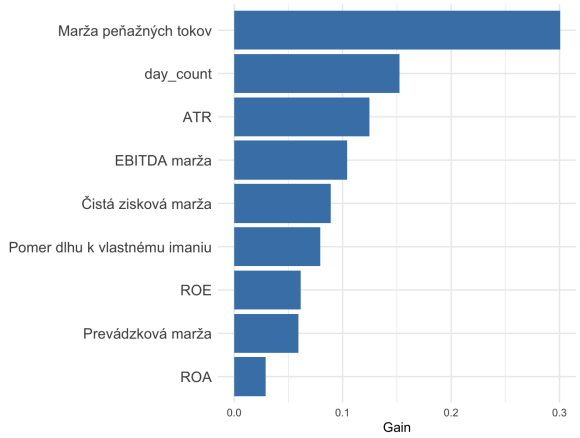
Obr. 11: Frekvencia prediktorov pre XGB, IQR filtrácia, premenná ln $ev2a$.



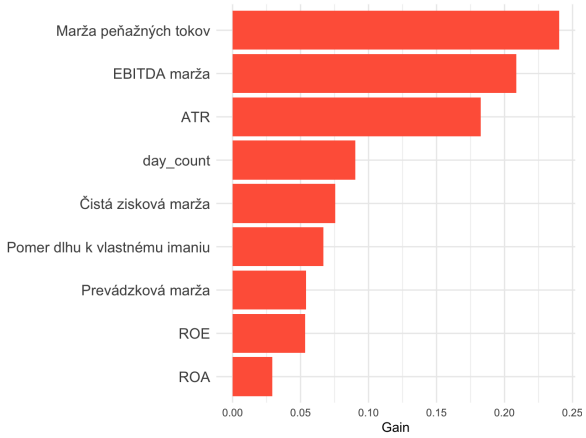
Obr. 12: Frekvencia prediktorov pre XGB, PCA filtrácia, premenná ln $ev2a$.

V prípade cieľovej premennej ln $ev2s$, na základe obrázkov Obr. 13, Obr. 14, Obr. 15, Obr. 16, Obr. 17 a Obr. 18, dosahuje najvyššiu hodnotu zlepšenia marža peňažných tokov, čo predstavuje rozdiel oproti cieľovým premenným ln $m2b$ a ln $ev2a$. V metrikách pokrytia a frekvencie dosahujú najvyššie hodnoty ROE a pomer dlhu k vlastnému imaniu. Po aplikácii PCA filtrovania sa v zlepšení výraznejšie uplatňuje aj EBITDA marža.

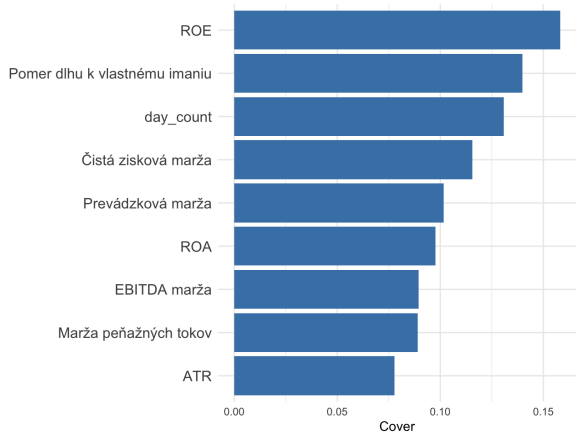
Výsledky analýz ukazujú, že medzi najvýznamnejšie prediktory naprieč všetkými cieľovými premennými patria ukazovatele kapitálovej štruktúry (pomer dlhu k vlastnému imaniu) a rentability (ROE, ROA). Významnú úlohu zohrávali aj ATR a day_count, ktoré sa opakovane umiestňovali vysoko vo všetkých troch metrikách. Naopak, maržové



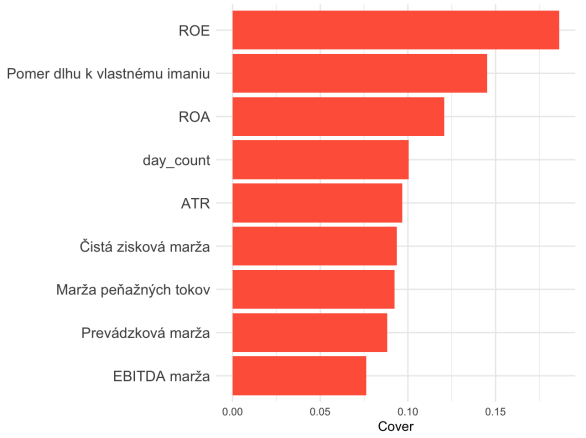
Obr. 13: Zlepšenie pre XGB podľa prediktorov, IQR filtrácia, premenná ln $ev2s$.



Obr. 14: Zlepšenie pre XGB podľa prediktorov, PCA filtrácia, premenná ln $ev2s$.



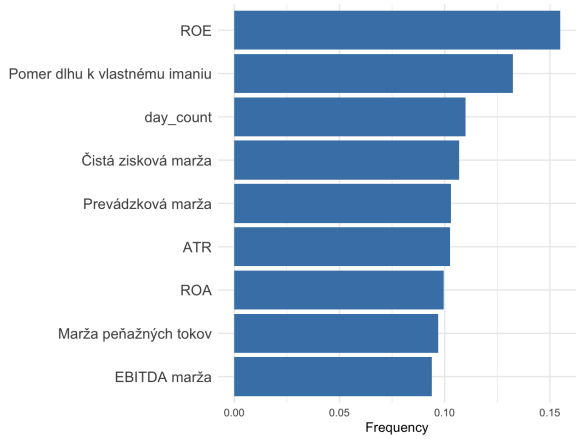
Obr. 15: Pokrytie prediktorov pre XGB, IQR filtrácia, premenná ln $ev2s$.



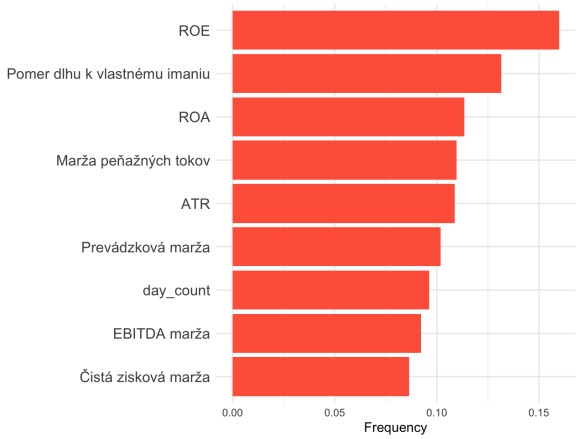
Obr. 16: Pokrytie prediktorov pre XGB, PCA filtrácia, premenná ln $ev2s$.

ukazovatele sa vo väčšine prípadov nachádzali na nižších priečkach a ich prínos k presnosti modelu bol menej výrazný. Výnimkou bola predikcia ln $ev2s$, kde sa ako dôležité ukázali marža peňažných tokov a $EBITDA$ marža.

Doposiaľ sme sa pri modeli XGBoost zameriavali na analýzu významnosti prediktorov, avšak stále zostáva nezodpovedaná otázka, akým spôsobom jednotlivé prediktory ovplyvňujú výsledné predikcie. Tieto úvahy sú obzvlášť dôležité z hľadiska analýzy podľa teórie hodnoty: predikčná presnosť síce predstavuje kľúčovú metriku, no zároveň je žiaduce, aby model správne vystihoval vzťahy medzi prediktormi a cieľovými premennými, ktoré vyplývajú z rovníc (31), (33) a (35). Na skúmanie čiastkových vplyvov prediktorov na výsledné

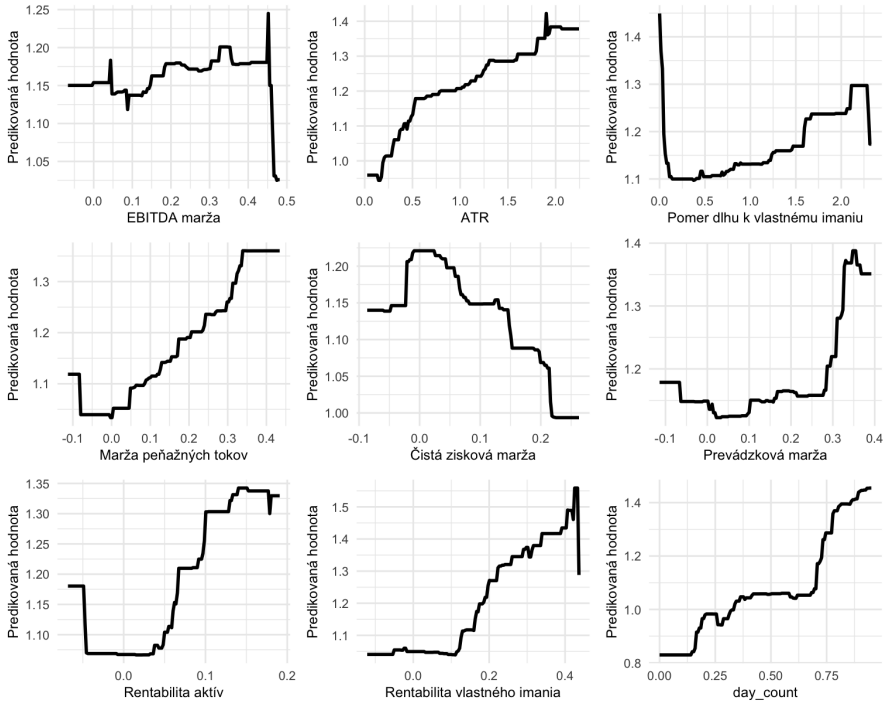


Obr. 17: Frekvencia prediktorov pre XGB, IQR filtrácia, premenná ln $ev2s$.



Obr. 18: Frekvencia prediktorov pre XGB, PCA filtrácia, premenná ln $ev2s$.

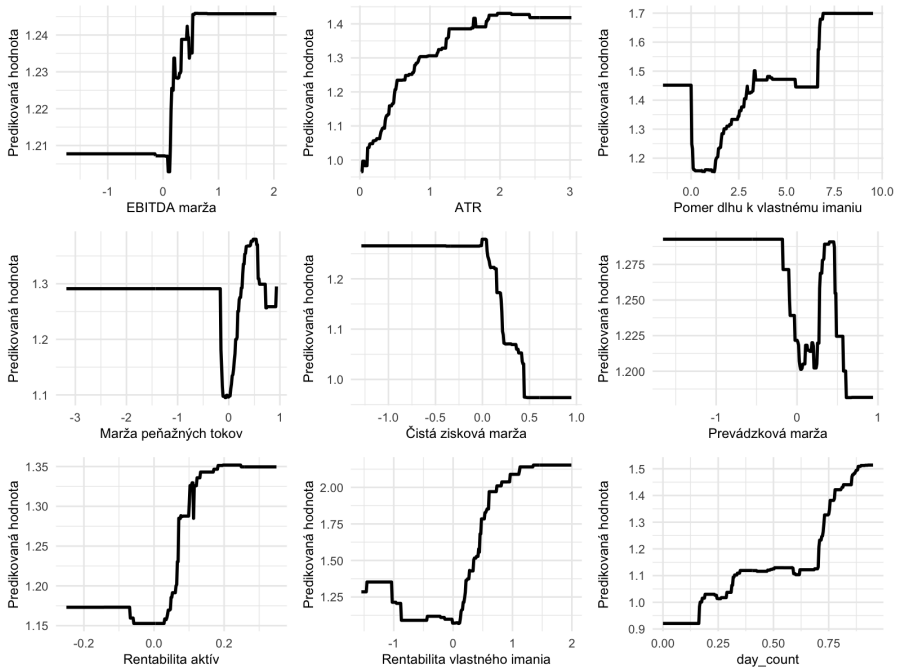
predikcie použijeme tzv. *partial dependence* grafy. Ich praktickú implementáciu popisuje článok [27]



Obr. 19: *Partial dependence plots* pre IQR filtrovanie a ln $m2b$.

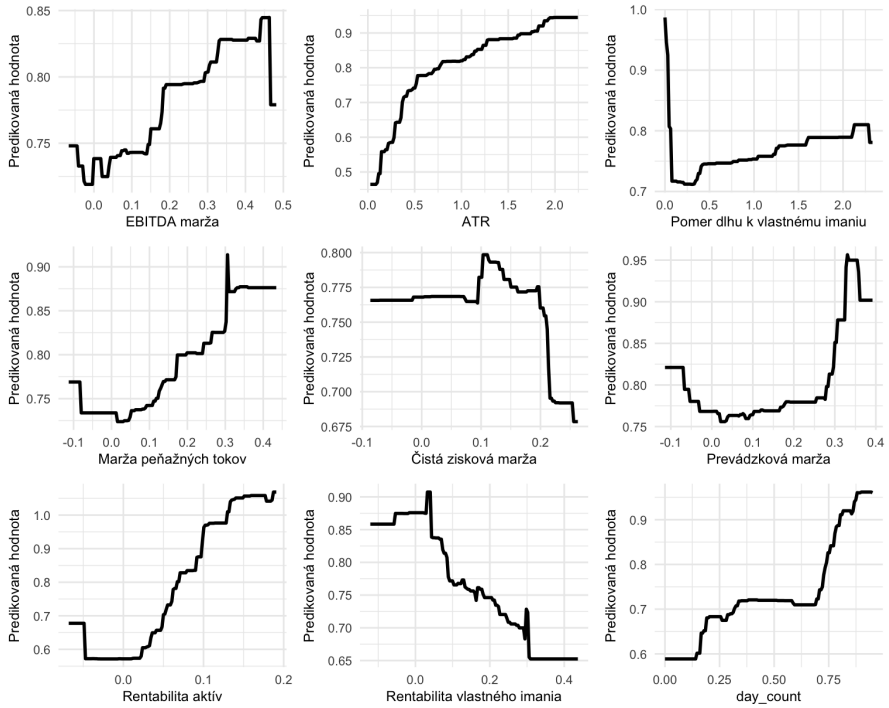
Poznámka 3.2.6. Keďže logaritmus je rastúca transformácia, pri interpretácii vzťahov medzi účtovnými ukazovateľmi a násobkom v rovnici (31) nezáleží na tom, či celú rovnicu (31) zlogaritmujeme. Čiastkový vplyv účtovných ukazovateľov zostáva kvalitatívne

rovnaký. Rovnako uvažujeme aj pre $ev2a$ a $ev2s$.

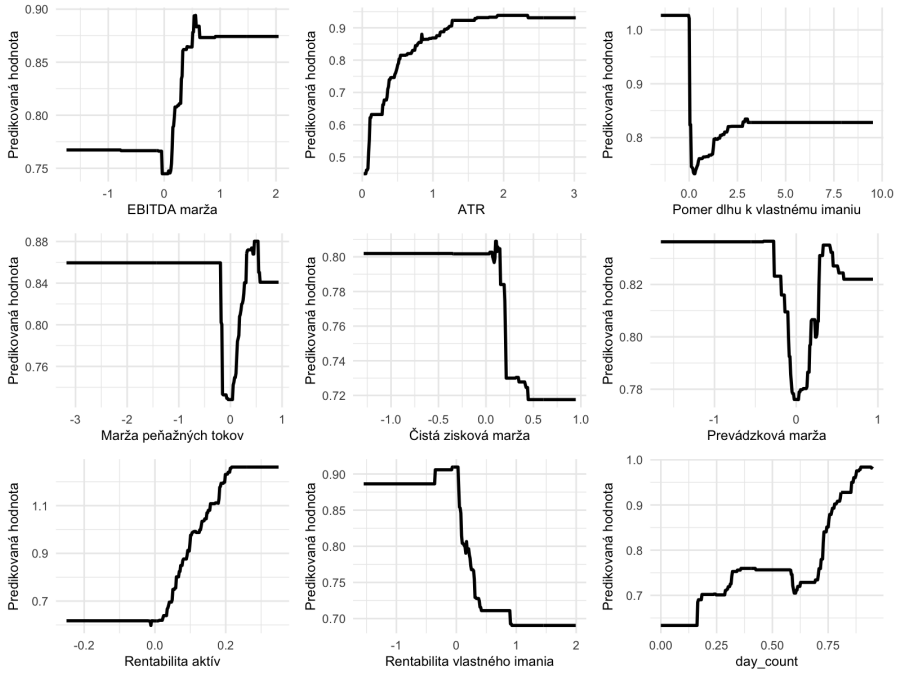


Obr. 20: *Partial dependence plots* pre PCA filtrovanie a ln $m2b$.

Z porovnania *partial dependence* grafov na Obr. 19 a Obr. 20 pre $\ln m2b$ vyplýva, že pre obe filtrovania možno pozorovať výrazný pozitívny vplyv ROE, čo je v súlade s analytickým vyjadrením $m2b$ v rovnici (31), kde je $m2b$ priamo úmerné ROE. Tento vzťah je najvýraznejší pri PCA filtrovaní, kde pri vysokých hodnotách rentability dochádza k prudkému nárastu predikovanej hodnoty. Prediktor *day_count* má v oboch filtrovaniach rastúci efekt, čo opäť naznačuje, že násobky v čase rastú. Naopak, niektoré maržové prediktory ako čistá zisková marža a prevádzková marža vykazujú protichodné správanie, čo môže naznačovať ich obmedzenú schopnosť vysvetliť variabilitu $m2b$, čo sme naznačili aj v časti popisujúcej významnosť prediktorov.



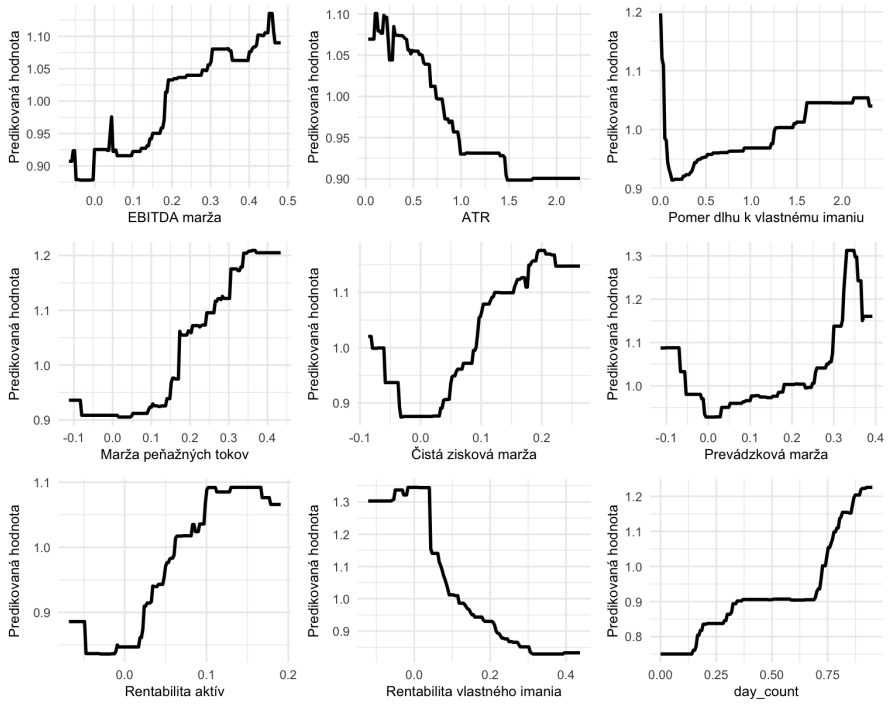
Obr. 21: *Partial dependence plots* pre IQR filtrovanie a ln $ev2a$.



Obr. 22: *Partial dependence plots* pre PCA filtrovanie a ln $ev2a$.

Porovnaním Obr. 21 a Obr. 22 získavame pohľad na vplyv prediktorov na predikované hodnoty $\ln ev2a$. Výsledky ukazujú konzistentný rast predikcií pri zvyšovaní ROA,

čo zodpovedá teoretickému očakávaniu podľa rovnice (33), kde $ev2a$ priamo závisí od ukazovateľa ROA. Pozitívny vplyv vykazujú aj ukazovatele ako EBITDA marža či marža peňažných tokov. Hoci tieto premenné nie sú súčasťou analytického výrazu, ich rast typicky signalizuje zlepšenie ziskovosti firmy, čo by malo zvýšiť násobok $ev2a$.



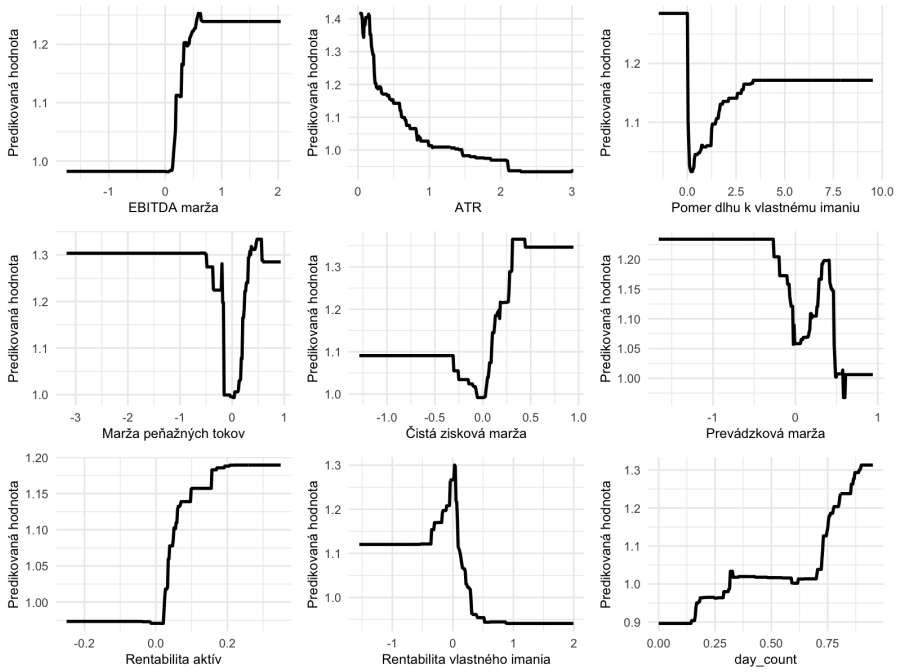
Obr. 23: *Partial dependence plots* pre IQR filtrovanie a ln $ev2s$.

Poslednými porovnávanými sú Obr. 23 a Obr. 24 pre cieľovú premennú ln $ev2s$. Premenná $ev2s$ závisí od ROA, ktorá má v oboch filtrovaniach silne rastúci vplyv. ATR, ktorý sa vo vzorci (35) nachádza v menovateli, má klesajúci priebeh, najmä pri PCA, čo je v súlade s teóriou.

Z výsledkov analýzy *partial dependence* grafov vyplýva, že smer vplyvu kľúčových prediktorov je vo väčšine prípadov konzistentný s analytickými vzťahmi (31), (33) a (35). Predovšetkým ROA a ROE majú stabilný pozitívny vplyv na cieľové premenné, zatiaľ čo ATR sa pri $ev2s$ správne prejavuje ako negatívny faktor. Rozdiely medzi filtráciami sa prejavujú najmä v rozsahu extrémnych hodnôt, pričom PCA zachytáva širšie spektrum variability.

Čo sa týka interpretácie SVR, vo všeobecnosti sa transformáciou cez kernelovú funkciu strácajú informácie. Tak ako je uvedené aj v článku [57], väčšina autorov od interpretácie

upúšťa. Vysvetlenie a vizualizácie podobné ako pre prípad XGB sú možné len za veľmi špecifických podmienok, ktoré predkladaná analýza nespĺňa. Z toho dôvodu sme interpretáciu a vizualizáciu nezrealizovali.



Obr. 24: *Partial dependence plots* pre PCA filtrovanie a ln $ev2s$.

4 Riadenie portfólia

V tejto kapitole sa zaoberáme problematikou riadenia investičného portfólia. Najskôr stručne predstavíme a následne empiricky overíme tri odlišné investičné stratégie: naivne diverzifikované portfólio, portfólio optimalizované podľa Markowitzovho modelu a vlastné LONG-SHORT portfólio založené na predikcii pomocou účtovných ukazovateľov (využívajúc automatizovaný model oceňovania firiem z prechádzajúcej časti). Prvé dve stratégie sú tradične považované za benchmark, pričom slúžia na porovnanie efektívnosti novonavrhnutých prístupov. Tretia stratégia bude predmetom detailnejšej analýzy, keďže predstavuje príspevok autora predkladanej práce.

Pri nákupe akcií investori očakávajú buď rast alebo pokles ceny. Ak predpokladajú rast ceny, zaujímajú tzv. LONG pozíciu. Investor vstupujúci do SHORT pozície predáva finančné aktívum, ktoré sám nevlastní, s cieľom odkúpiť ho neskôr za nižšiu cenu. Takéto aktívum si predajca vypožičia, spravidla od brokera alebo inej finančnej inštitúcie, pričom sa zaväzuje ho v budúcnosti vrátiť. Za poskytnutie aktíva platí poplatok nazývaný *borrowing fee*, ktorý predstavuje formu nájomného za zapožičanie. Okrem toho si broker vypýta aj likvidnú zábezpeku od 25% do 50% absolútnej hodnoty SHORT pozície [36]. Ekonomickou motiváciou tejto stratégie je predpoklad poklesu trhovej ceny daného aktíva, čo umožní investorovi dosiahnuť zisk z cenového rozdielu medzi predajom a následným spätným nákupom. Pri výpočtoch v tejto práci abstrahujeme od transakčných nákladov, neberieme do úvahy *borrowing fee* a predpokladáme, že aktíva sú dokonale deliteľné a obchodovateľné bez obmedzení.

Definícia 4.0.1. Nech $t = 0, \ldots$ označuje časovú periódu, nech $i = 1, \ldots, D$ je index akcie. Pod výnosom akcie i v časovej perióde $t + 1$ rozumieme

$$r_{t+1,i} = \frac{P_{t+1,i}}{P_{t,i}} - 1, \tag{70}$$

kde $P_{t,i}$ je cena akcie i v perióde t. V zmysle definície 4.0.1 definujeme ex-post výnos portfólia ako

$$R_{t+1} = \sum_{i=1}^{D} w_{t,i} \times r_{t+1,i}, \tag{71}$$

kde $w_{t,i}$ je váha akcie i v portfóliu na začiatku periódy, celý vektor váh značíme $\mathbf{w}_t = (w_{t,1}, \ldots, w_{t,D}) = (w_1, \ldots, w_D)_t$. Časový index t volíme vo vzorci (71) pre $w_{t,i}$ schválne,

pretože medzi časovými periódami t a $t+1$ nerebalansujeme. Horným indexom budeme pri všetkých symboloch značiť portfóliový prístup. Symbol $\mathbf{r}_{t+1}$ reprezentuje stĺpcový vektor výnosov. Číslo D je počet akcií v indexe S&P500, v tejto práci $D = 497$.

Poznámka 4.0.2. Je zrejmé, že v prípade SHORT pozície je príslušná váha záporná.

Na vyhodnotenie výkonnosti a rizikovosti portfóliových prístupov budeme využívať ešte dve metriky. Prvou z nich je hrubá expozícia, ktorú rátame ako

$$HE_t = \sum_{i=1}^{D} |w_{t,i}|. \tag{72}$$

Táto metrika ukazuje, ako veľmi je stratégia agresívna a využíva finančnú páku. Zrejme platí, že ak $HE_t \leq 1$, portfóliová stratégia nevyužíva žiadne dodatočné úvery alebo páku, a teda riziková angažovanosť je obmedzená na výšku vlastného kapitálu. Alternatívne, ak $HE_t > 1$, portfólio je vystavené väčšiemu riziku. Druhou dôležitou metrikou je čistá expozícia v čase t, ktorú rátame ako

$$NE_t = \sum_{i=1}^{D} w_{t,i}. \tag{73}$$

Interpretácia je zrejmá: v prípade, že je $NE_t = 0$, portfólio je neutrálne (vyvažuje LONG a SHORT pozície). Ak je $NE_t > 0$, portfóliový prístup je nastavený na rast celého trhu. V prípade $NE_t < 0$ je orientované na pokles cien akcií.

4.1 Naivne diverzifikované portfólio (N)

Naivne diverzifikované portfólio predstavuje investičnú stratégiu, kde sa investície rovnomerne rozdeľujú medzi D dostupných akcií. Neprihliadame pritom na ich individuálne charakteristiky. Toto portfólio je ekvivalentné s LONG pozíciou pre všetky dostupné akcie. Stratégiu sme realizovali týmto spôsobom

- Zaujali sme LONG pozíciu pre každú z D akcií k 31.12.2023.
- Na konci obdobia, t.j. 31.12.2024, sme zrátali ex-post výnos pre celé portfólio.

Možno teda napísať, že

$$\left(\mathbf{w}_{2023}^{N}\right)^{\mathsf{T}} = \overbrace{\left(\tfrac{1}{D}, \ldots, \tfrac{1}{D}\right)_t^{\mathsf{T}}}^{D\text{-krát}}.$$

Dosiahnutý ex-post výnos stratégie bol $R_{2024}^{N} = \left(\mathbf{w}_{2023}^{N}\right)^{\mathsf{T}} \mathbf{r}_{2024} = 15{,}73\%$. Zrejme platí $NE_{2023} = HE_{2023} = 1$.

4.2 Markowitzovo portfólio

V tejto podkapitole v krátkosti popíšeme a prakticky implementujeme dva varianty Markowitzovho portfólia - tzv. *global minimum variance* a *mean-variance*. Myšlienkami aj kódom sledujeme rovnakú logiku ako autori v [50] a [65].

4.2.1 *Global minimum variance* portfólio (G)

Global minimum variance predstavuje portfólio s najnižšou možnou varianciou výnosu spomedzi všetkých dostupných portfólií zostaviteľných z D akcií. [65]. Nezohľadňuje očakávaný výnos. Optimalizačný problém má tvar

$$\min_{\mathbf{w_t}^G} \quad \left(\mathbf{w_t}^G\right)^\top \mathbf{\Sigma} \left(\mathbf{w_t}^G\right), \tag{74}$$

$$\text{s podmienkou} \quad \mathbf{1}^\top \left(\mathbf{w_t}^G\right) = 1, \tag{75}$$

kde $\left(\mathbf{w_t}^G\right)^\top = (w_1^G, \ldots, w_D^G)_t^\top$ je vektor váh akcií a $\mathbf{\Sigma} \in \mathbb{R}^{D \times D}$ je kovariančná matica výnosov D aktív. Riešenie optimalizačného problému sa dá zapísať ako

$$\mathbf{w}_t^G = \frac{\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^\top \mathbf{\Sigma}^{-1}\mathbf{1}}, \tag{76}$$

kde $\mathbf{\Sigma}^{-1}$ je inverzná matica k matici $\mathbf{\Sigma}$ a $\mathbf{1}$ je jednotkový vektor. G portfólio tvorí začiatočný bod efektívnej hranice z Markowitzovej teórie [65] a slúži ako referenčný model pri porovnávaní portfólií založených na výnose alebo averzii k riziku. Stratégiu sme realizovali nasledujúcim spôsobom:

- Stiahli sme denné upravené zatváracie ceny všetkých D akcií za obdobie od 31.12.2022 až 31.12.2023.
- Na základe zatváracích dát sme spočítali denné výnosy D akcií a zrátali sme $\mathbf{\Sigma}_{day}$ ako kovariančnú maticu denných výnosov, ktorú sme anualizovali (t.j. prenásobili počtom obchodovacích dní (252)), aby sme získali $\mathbf{\Sigma}$. Následne sme vypočítali $\mathbf{\Sigma}^{-1}$.
- Vyrátali sme $\mathbf{w}_{2023}^G$.
- Zaujali sme LONG, resp. SHORT pozíciu pre každú z D k 31.12.2023 podľa $\mathbf{w}_{2023}^G$.

- Na konci obdobia, t.j. 31.12.2024 sme zrátali ex-post výnos pre celé portfólio.

Pre detailnejší postup viď. kód v Príloha B: Markowitzovo portfólio. Ex-post výnos $R^G_{2024} = \left(\mathbf{w}^G_{2023}\right)^\mathsf{T} \mathbf{r}_{2024} = 20{,}44\%$. Čo sa týka rizikového profilu G portfólia, HE_{2023} je až na úrovni 3,2, čo značí vyššiu dlhovú angažovanosť portfólia ako v prípade N portfólia. Zrejme $NE_{2023} = 1$, keďže je to v samotných podmienkach optimalizačnej úlohy.

4.2.2 Portólio optimalizované cez Sharpeov pomer (S)

Je to portfólio, ktoré maximalizuje Sharpeov pomer, t. j. pomer medzi rizikovo upraveným výnosom a smerodajnou odchýlkou portfólia

$$\text{Sharpeov pomer} = \frac{\mathbb{E}[R_p] - r_f}{\sigma_p}, \tag{77}$$

kde $\mathbb{E}[R_p]$ je očakávaný výnos portfólia, r_f je bezriziková miera a σ_p je smerodajná odchýlka portfólia. Je nutné poznamenať, že σ_p a $\mathbb{E}[R_p]$ sú neznáme a odhadujeme ich veličinami v (78). Cieľom je nájsť také váhy $\mathbf{w}^S_t = (w^S_1, \ldots, w^S_D)^\top_t$, ktoré riešia optimalizačný problém

$$\max_{\mathbf{w}^S_t} \quad \frac{\left(\mathbf{w}^S_t\right)^\top (\boldsymbol{\mu} - r_f \mathbf{1})}{\sqrt{\left(\mathbf{w}^S_t\right)^\top \boldsymbol{\Sigma} \left(\mathbf{w}^S_t\right)^\top}}, \tag{78}$$

$$\text{s podmienkou} \quad \left(\mathbf{w}^S_t\right)^\top \mathbf{1} = 1. \tag{79}$$

kde $\boldsymbol{\mu} \in \mathbb{R}^D$ je vektor očakávaných výnosov D akcií odhadnutý ako anualizovaný priemer z historickej výkonnosti (t.j. priemer denných výnosov prenásobený počtom obchodovacích dní), $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$ je kovariančná matica historických výnosov a r_f je bezriziková miera. Tento problém možno riešiť v dvoch krokoch [50] . Najprv sa určí smerový vektor (neškálované portfólio) ako

$$\mathbf{w}^{\text{raw}} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1}). \tag{80}$$

Následne sa tento vektor normuje tak, aby súčet váh bol rovný jednej, čo možno zapísať ako

$$\left(\mathbf{w}_t^S\right)^\top = \frac{\mathbf{w}^{\text{raw}}}{\mathbf{1}^\top \mathbf{w}^{\text{raw}}}, \tag{81}$$

kde $\left(\mathbf{w}_t^S\right)^\top$ sú finálne investičné váhy S portfólia. Stratégiu sme realizovali nasledujúcim spôsobom:

- Stiahli sme denné upravené zatváracie ceny všetkých D akcií za obdobie od 31.12.2022 až 31.12.2023.
- Na základe historických dát sme odhadli $\boldsymbol{\Sigma}$ (ako v G portfóliu) a $\boldsymbol{\mu}$.
- Odhadli sme bezrizikovú mieru na základe výnosu 10 ročného amerického dlhopisu ([58]), ktorý bol 0,03898.
- Zaujali sme LONG, resp. SHORT pozíciu pre každú z D akcií k 31.12.2023 podľa váh $\mathbf{w}_{2023}^S$.
- Na konci obdobia, t.j. 31.12.2024 sme zrátali ex-post výnos pre celé portfólio.

Pre detailnejší postup viď. kód v Príloha B: Markowitzovo portfólio. Ex-post výnos tejto stratégie bol $R_{2024}^S = \left(\mathbf{w}_{2023}^S\right)^\mathsf{T} \mathbf{r}_{2024} = 27{,}4\%$, no očakávaný výnos $\boldsymbol{\mu}^\top \mathbf{w}_{2023}$ bol 396%. To poukazuje na fakt, že očakávané výnosy odhadnuté z historických dát negarantujú aj budúce výnosy. Okrem toho $HE_{2023} = 9{,}76$, čo naznačuje ešte väčšiu dlhovú zaangažovanosť ako v prípade G portfólia. Zrejme $NE_{2023} = 1$.

4.3 LONG-SHORT portfólio (LS)

V tejto časti práce prezentujeme implementáciu investičnej stratégie, ktorá využíva výstupy z predikčných modelov XGB a SVR. Základná myšlienka stratégie spočíva v snahe identifikovať nesprávne ocenené akcie na základe rozdielu medzi predikciami a skutočnými hodnotami násobkov $m2b$, $ev2a$ a $ev2s$.

Definícia 4.3.1. Označme $\hat{y}_{t,i}$ predikovanú hodnotu cieľovej premennej pre akciu i v čase t a $y_{t,i}$ jej skutočnú hodnotu. Potom pod nesprávnym ocenením (tzv. mispricingom)

rozumieme číslo $M_{t,i}$ definové ako

$$M_{t,i} = \frac{\hat{y}_{t,i} - y_{t,i}}{y_{t,i}}. \tag{82}$$

Rovnica (82) vyjadruje signál odchýlky medzi predikovanou a skutočnou hodnotou cieľovej premennej, pričom vyššia absolútna hodnota indikuje silnejší nesúlad v relatívnom ocenení akcie. Na základe výsledkov uvedených v Tab. 2 a Tab. 3 možno konštatovať, že mediánová relatívna odchýlka sa pohybuje v intervale od 0,254 do 0,345. Pri konštrukcii LONG-SHORT stratégie sme preto pozornosť zamerali na výraznejšie signály s vyššou pravdepodobnosťou správneho určenia smeru odchýlky. Zavádzame preto prah τ, ktorému sme stanovili hodnotu 0,345 (na úrovni maximálnej mediánovej relatívnej chyby zo všetkých nelineárnych modelov). Pomocou τ model rozhodol, pre ktoré signály máme v portfóliu zaujať nejakú pozíciu, čo je ilustrované v nasledujúcej definícii.

Definícia 4.3.2. Nech $i = 1, \ldots, D$ označuje index akcie, t nech je časová perióda. Rozhodnutie o pozícii pri akcii i v čase t prebieha podľa pravidla

$$\text{pozícia}_{t,i} = \begin{cases} \text{LONG} & \text{ak } M_{t,i} \geq \tau, \\ \text{SHORT} & \text{ak } M_{t,i} \leq -\tau, \\ \text{NEUTRÁLNA} & \text{ak } M_{t,i} \in (-\tau, \tau), \end{cases} \tag{83}$$

kde NEUTRÁLNA znamená, že nezaujmeme ani pozíciu LONG ani SHORT, teda váha v portfóliu bude 0.

Po určení typu pozície je potrebné rozhodnúť, aký podiel celkového kapitálu bude investovaný do jednotlivých akcií. Túto úlohu riešime pomocou tzv. *confidence*. V tomto kontexte ide o mieru presvedčenia modelu, že nesprávne ocenenie je dostatočne výrazné.

Definícia 4.3.3. Nech má akcia $i = 1, \ldots, D$ v čase t mispricing $M_{t,i}$. Potom *confidence* definujeme ako

$$C_{t,i} = \frac{\max(0, |M_{t,i}| - \tau)}{\sum_{j=1}^{D} \max(0, |M_{t,i}| - \tau)}. \tag{84}$$

Predpokladáme, že existuje také j pre každé t, že $C_{t,j} \neq 0$. Pre vektor $\mathbf{w}_t^{LS}$ platí

$$\left(\mathbf{w}_t^{LS}\right)^{\mathsf{T}} = \left(w_1^{LS}, \ldots, w_D^{LS}\right)_t^{\top} = \left(C_{t,1}, \ldots, C_{t,D}\right)^{\top}.$$

Poznámka 4.3.4. Dá sa nahliadnuť, že aj váhy pre pozície SHORT sú v LONG-SHORT prístupe kladné. Túto nekonzistentnosť vo výpočtoch sme odstránili tak, že sme výnos rátali inak ako v 4.0.1. Definícia výnosu v prípade LONG-SHORT portfólia je potom

$$r_{t+1,i} = \begin{cases} \frac{P_{t+1,i}}{P_{t,i}} - 1 & \text{ak pozícia}_{t,i} \text{ je LONG,} \\ 1 - \frac{P_{t+1,i}}{P_{t,i}} & \text{ak pozícia}_{t,i} \text{ je SHORT.} \end{cases} \tag{85}$$

Okrem toho je nutné mierne upraviť aj definíciu NE_t ako

$$NE_t = \sum_{i:\text{LONG}}^{D} w_{t,i} - \sum_{i:\text{SHORT}}^{D} w_{t,i}. \tag{86}$$

Takto zavedená LONG-SHORT stratégia má niekoľko praktických implikácií. Prvou je, že hrubá expozícia celého portfólia je vždy 1. To značí, že celá stratégia je krytá dostupným kapitálom a je konzervatívnejšia ako prístupy prezentované v časti o Markowitzovom portfóliu.

Keďže máme niekoľko cieľových premenných, výsledky prezentujeme v prehľadných Tab. 6, Tab. 7 a Tab. 8. V poslednej menovanej tabuľke ukazujeme aj NE_{2023}. Čistá expozícia NE_{2023} ukazuje, že modely majú tendenciu preferovať LONG pozície naprieč všetkými cieľovými premennými a filtráciami. Z regresných modelov vykazuje SVM konzistentne najvyššiu čistú expozíciu. Zároveň sa ukazuje, že väčšinou je vyššia čistá expozícia pri PCA filtrácii.

Model	**Cieľ. premenná**	**Výnos LONG %**	**Výnos SHORT %**	**Výnos port. [%]**
SVM	$ev2a$	20,1	1,3	21,4
SVM	$ev2s$	24,9	1,4	26,3
SVM	$m2b$	21,4	−1,3	20,1
XGB	$ev2a$	19,1	0,8	19,8
XGB	$ev2s$	22,4	3,2	25,6
XGB	$m2b$	28,4	−4,3	24,1

Tab. 6: Ex-post výnosy pre LONG a SHORT časti portfólia pre jednotlivé filtrovania a cieľové premenné pre IQR filtrovanie.

Model	Cieľ. premenná	Výnos LONG %	Výnos SHORT %	Výnos port. [%]
SVM	$ev2a$	20,8	0,8	21,6
SVM	$ev2s$	23,8	1,3	25,1
SVM	$m2b$	18,2	−1,9	16,3
XGB	$ev2a$	24,3	1,8	26,1
XGB	$ev2s$	28,1	0,9	29,0
XGB	$m2b$	16,3	−1,4	14,9

Tab. 7: Ex-post výnosy pre LONG a SHORT časti portfólia pre jednotlivé filtrovania a cieľové premenné pre PCA filtrovanie.

Filtrácia	Model	Cieľová premenná	LONG %	SHORT %	NE_{2023} %
IQR	SVM	$ev2a$	88,4	11,6	76,8
IQR	SVM	$ev2s$	90	10	80
IQR	SVM	$m2b$	91,2	8,8	82,4
IQR	XGB	$ev2a$	77,2	22,8	54,4
IQR	XGB	$ev2s$	76,6	23,4	53,2
IQR	XGB	$m2b$	82	18	64
PCA	SVM	$ev2a$	89,5	10,5	79
PCA	SVM	$ev2s$	89,6	10,4	79,2
PCA	SVM	$m2b$	92,2	7,8	84,4
PCA	XGB	$ev2a$	81,9	18,1	63,8
PCA	XGB	$ev2s$	76,7	23,3	53,4
PCA	XGB	$m2b$	90,1	9,9	80,2

Tab. 8: Percentuálna veľkosť LONG a SHORT pozícií portfólia, čistá expozícia portfólia.

V tabuľkách Tab. 6 a Tab. 7 sú výnosy portfólií rozdelené podľa typu filtrácie a cieľovej premennej. Z nich vyplýva, že LONG časť prináša vždy dvojciferné výnosy, zatiaľ čo SHORT sa väčšinou drží na jednotkách percent a pri $m2b$ je dokonca záporná. Pri IQR filtrovaní dosiahol najvyšší portfóliový výnos SVM model na $ev2s$ vo výške $26,3\%$, XGB na tej istej premennej zaznamenal $25,6\%$ a zároveň vygeneroval najziskovejší SHORT

vo výške $+3,2\%$. Po aplikovaní PCA filtrovania u XGB modelu sa LONG výnosy pre premenné $ev2a$ a $ev2s$ podstatne zvýšili – na $24,3\%$ a $28,1\%$, čo viedlo k portfóliovému výnosu $29,0\%$. Naopak pri $m2b$ PCA celkový výnos znížilo. Z týchto výsledkov vyplýva, že pri cieľových premenných $ev2a$ a $ev2s$ dokáže PCA vylepšiť LONG pozície, no pri $m2b$ je trend opačný. SVM zasa vykazuje porovnateľné výsledky pri oboch filtráciách, s výnimkou prípadu $m2b$, kde pri PCA nastal pokles z 20,1% na 16,3%.

Záver

V predkladanej diplomovej práci sme sa zamerali na zostrojenie automatizovaného modelu oceňovania verejne obchodovaných spoločností. Úvodná kapitola poskytla prehľad a porovnanie oceňovacích prístupov s cieľom identifikovať vhodný rámec pre jeho konštrukciu. Hoci boli predstavené aj princípy diskontovania peňažných tokov, ukázali sme, že relatívne oceňovanie predstavuje vhodnejší základ pre automatizáciu — najmä preto, že vychádza z trhových dát, eliminuje subjektívne predpoklady analytika a je metodicky jednoduchšie a priamočiarejšie. Za veľmi prínosné považujeme uvedenie rovníc (31), (33) a (35), ktoré ilustrujú prepojenie medzi teóriou hodnoty a konceptom relatívneho oceňovania. Tieto analytické vzťahy zároveň poskytujú teoretický rámec, voči ktorému sme porovnávali výstupy modelov strojového učenia s cieľom overiť, či ich predikcie zodpovedajú očakávaniam vyplývajúcim z fundamentálnej teórie.

V druhej kapitole sme stručne predstavili regresné techniky, ktoré sme využili pri konštrukcii automatizovaného modelu oceňovania firiem. Pre support vector regresiu sme vychádzali z článkov [6], [7] a [1], v prípade lineárnej regresie z práce [34]. Pri XGBoost modeli sme sa opierali priamo o článok autorov, ktorí tento algoritmus navrhli [33].

V tretej kapitole sme najprv odôvodnili výber použitých dát a následne podrobne popísali celý proces ich predspracovania. Detailný popis slúži ako návod pre ďalších autorov, aby sa vyhli potenciálnym úskaliam, na ktoré sme počas práce narazili. Pri identifikácii odľahlých pozorovaní sme otestovali dva prístupy. Prvým bol klasický postup založený na IQR, ktorý však viedol k výraznej redukcii objemu dát. Preto sme sa rozhodli pre netradičné využitie založené na princípoch PCA. Ako ukazujú výsledky v Tab. 2 a Tab. 3, zvolený prístup viedol k zlepšeniu výkonnosti modelov, najmä z pohľadu metriky R^2. Okrem toho výsledky naznačujú, že lineárna regresia dosahuje nižšiu predikčnú presnosť v porovnaní s modernejšími metódami, ktoré sme v práci ďalej analyzovali.

Okrem lineárneho modelu sme sa zamerali aj na pokročilé metódy interpretácie výstupov XGBoost modelu. Za mimoriadne prínosné považujeme Obr. 19, Obr. 20, Obr. 21, Obr. 22, Obr. 23 a Obr. 24, ktoré pomáhajú lepšie pochopiť správanie modelu XGBoost a jeho predikcie. V rovnakej časti práce sú uvedené aj vizualizácie významnosti prediktorov, vyjadrené ich vplyvom na rozdelenie uzlov v rámci stromovej štruktúry. V tejto časti sme zároveň posudzovali, či kvalitatívne vzťahy medzi prediktormi a cieľovými premennými

korešpondujú s očakávaniami vyplývajúcimi z teórie hodnoty.

V prípade SVR sme obdobnú analýzu neuskutočnili, keďže v prostredí R a v rámci balíka pdp nie je dostupná funkcia, ktorá by umožňovala vytváranie PDP grafov pre tento typ modelu. V odbornej literatúre sa nám podarilo identifikovať iba jediný článok [57], ktorý sa venuje interpretovateľnosti SVR modelov. Aj samotní autori v ňom uvádzajú, že vizualizácia je možná len pri špecifickej štruktúre vstupných dát a pri použití konkrétne definovanej kernelovej funkcie.

V poslednej kapitole sme sa zamerali na praktickú implementáciu viacerých prístupov k riadeniu portfólia. Ako východiskový prístup sme zvolili naivne diverzifikované portfólio. Následne sme aplikovali modely založené na Markowitzovej teórii, konkrétne portfólio s minimálnou varianciou (GMV) a portfólio maximalizujúce Sharpeov pomer. Upozornili sme pritom na citlivosť týchto modelov voči vstupným parametrom — hoci očakávaný výnos bol vysoký, ex-post výnos bol výrazne nižší. Analýzu sme rozšírili aj o dve rizikové metriky (hrubú a čistú expozíciu), ktoré poskytujú informáciu o celkovej alokácii kapitálu a miere financovania portfólia dlhom, čím umožňujú lepšie pochopiť jeho rizikový profil.

Ako posledný sme opísali autorský prístup k riadeniu portfólia, konkrétne stratégiu typu LONG-SHORT. Investičná logika je založená na identifikácii nadhodnotených a podhodnotených akcií pomocou modelov strojového učenia trénovaných na trhových dátach (účtovných ukazovateľoch). V rámci návrhu sme zaviedli parameter τ, ktorý určuje prah pre výber dostatočne silných *mispricingov*. Alokácia kapitálu prebieha automatizovane na základe miery dôvery modelu (*confidence*)v danú predikciu, ktorú sme definovali v (84). Ex-post testovanie ukázalo, že tento prístup prináša výnosy porovnateľné s Markowitzovými modelmi, avšak pri nižšej dlhovej angažovanosti, o čom sa dá presvedčiť v Tab. 6 a Tab. 7.

Možnosti rozvoja tohto konceptu konštrukcie automatizovaných modelov sú stále otvorené. Veríme však, že práca poukázala na otázky, ktoré si zaslúžia ďalšiu pozornosť. Za osobitne perspektívnu považujeme tému kvantifikácie rizika v LONG-SHORT portfóliách. Okrem základných metrík ako čistá a hrubá expozícia sme sa tejto oblasti venovali len okrajovo. Ďalší výskum by mohol smerovať aj k optimalizácii výpočtového výkonu a ladeniu hyperparametrov použitých XGBoost modelom. Za zaujímavú oblasť považujeme aj snahu lepšie porozumieť fungovaniu SVR modelu pri predikciách. Hoci jeho výkonnosť je

porovnateľná s XGBoostom, ako ukazujú výsledky v Tab. 2 a Tab. 3, obmedzené možnosti interpretácie výstupov predstavujú nevýhodu z hľadiska praktickej využiteľnosti.

Zoznam použitej literatúry

[1] Awad, M., Khanna, R.: *Theories, Concepts, and Applications for Engineers and System Designers*, Apress, Berkeley, 2015

[2] Barrett, T. et al.: *data.table: Extension of 'data.frame'*, R manuál, 2025, dostupné na internete (4.5.2025): https://CRAN.R-project.org/package=data.table

[3] Bartram, S.M., Grinblatt, M.: *Agnostic fundamental analysis works*, Journal of Financial Economics 128 (2018), 125-147, dostupné na internete (4.5.2025): https://www.sciencedirect.com/science/article/abs/pii/S0304405X1730315X

[4] Bernoulli, D.: *Exposition of a New Theory on the Measurement of Risk*, Econometrica 22 (1954), 23-36, dostupné na internete (27.04.2025): http://69.164.214.130/cs/bc/dsci/Bernoulli_1738.pdf

[5] Bryan, J., Wickham H.: *readxl: Read Excel Files*, R manuál, 2025, dostupné na internete: https://CRAN.R-project.org/package=readxl

[6] Camps-Valls, G. et al.: *Support vector machines in engineering: an overview*, WIREs Data Mining and Knowledge Discovery 4, 234-267, dosupné na internete (31.1.2025): https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1125

[7] Cao, B., Zhan, D., Wu, X.: *Application of SVM in Financial Research*, International Joint Conference on Computational Sciences and Optimization (2009), príspevok na konferenciu, dostupné na internete (29.4.2025): https://ieeexplore.ieee.org/abstract/document/5194004

[8] Connel, J.: *Equity Valuation. Models from Leading Investment Banks*, John Wiley & Sons, Chichester, 2008

[9] Cornell, B.: *Medallion Fund: The Ultimate Counterexample?*, pracovný draft, The Anderson Graduate School of Management, 2020, dostupné na internete (5.5.2025): https://www.researchgate.net/profile/Bradford-Cornell/publication/338592203_Medallion_Fund_The_Ultimate_Counterexample/links/60914f2ea6fdccaebd08ef5d/Medallion-Fund-The-Ultimate-Counterexample.pdf

[10] Damodaran, A.: *Damodaran on Valuation. Security Analysis for Investment and Corporate Finance*, John Wiley & Sons, Hoboken, 2006

[11] Damodaran, A.: *Strategic Risk Taking: A Framework for Risk Management*, Pearson Education, Inc., Upper Saddle River, 2007

[12] Damodaran, A: *The Little Book of Valuation: How to Value a Company, Pick a Stock, and Profit*, John Wiley & Sons, Hoboken, 2011

[13] Damodaran, A: *Valuation Approaches and Metrics: A Survey of the Theory and Evidence*, Foundations and Trends® in Finance 1 (2007), 693-784, dostupné na internete (27.4.2025): https://www.nowpublishers.com/article/Details/FIN-013

[14] Damodaran, A.: *What is the riskfree rate? A Search for the Basic Building Block*, Journal of New Finance 1 (2020), dostupné na internete (27.4.2025): https://jnf.ufm.edu/cgi/viewcontent.cgi?article=1010&context=journal

[15] Dancho, M. Vaughan, D.: *tidyquant: Tidy Quantitative Financial Analysis*, R manuál, 2025, dostupné na internete: https://CRAN.R-project.org/package=tidyquant

[16] Deo, R.C.: *Machine Learning in Medicine*, Circulation 132 (2015), 1920-1930, dostupné na internete (5.5.2025): https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.115.001593

[17] Doğan, M. et al.: *Prediction of market value of firms with corporate sustainability performance data using machine learning models*, Finance Research Letters 77 (2025), dostupné na internete (4.5.2025): https://www.sciencedirect.com/science/article/abs/pii/S1544612325003484?dgcid=coauthor

[18] dos Santos, P.T.C. et al.: *Multiclass Legal Judgment Outcome Prediction for Consumer Lawsuits using XGBoost and TPE*, 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (2020), 881-886, dostupné na internete (29.4.2025): https://ieeexplore.ieee.org/document/9283094

[19] Ernst, D., Häcker, J.: Applied International Corporate Finance, Vahlen, Mníchov, 2011

[20] Fama, E.F., French, K.R.: *A five-factor asset pricing model*, Journal of Financial Economics 116 (2015), 1-22, dostupné na internete (27.4.2025): https://www.sciencedirect.com/science/article/abs/pii/S0304405X14002323

[21] Fatouros, G. et al.: *DeepVaR: a framework for portfolio risk assessment leveraging probabilistic deep neural networks*, Digital Finance 5 (2023), 29-56, dostupné na internete (29.4.2025): https://link.springer.com/article/10.1007/s42521-022-00050-0#citeas

[22] Financial modeling prep, zdroj dát, dostupné na internete: https://site.financialmodelingprep.com

[23] Friedman, J.H.: Greedy Function Approximation: A Gradient Boosting Machine, The Annals of Statistics 29 (2001), 1189-1232, dostupné na internete (29.4.2025): https://www.jstor.org/stable/2699986

[24] Gammerman, A. et al.: *Theory of Support Vector Machines*, učebný text, Royal Holloway University of London, Londýn, 1996, dostupné na internete (30.1.2025): https://ynucc.yu.ac.kr/~shkwon/lectures/ic/svm/svm_1.pdf

[25] Geertsema, P., Lu, H.: *Relative Valuation with Machine Learning, Journal of Accounting Research* 61 (2022), 329–376, dostupné na internete (18.1.2025): https://doi.org/10.1111/1475-679x.12464

[26] Graham, B.: *The Intelligent Investor*, HarperBusiness, New York, 2006

[27] Greenwell, B.M: *pdp: An R Package for Constructing Partial Dependence Plots pdp: An R Package for Constructing Partial Dependence Plots*, The R Journal 9 (2017) 420-436, dostupné na internete (2.5.2025): https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1600&context=r-journal

[28] Gupta,V.: *Predicting Accuracy of Valuation Multiples Using Value Drivers: Evidence from Indian Listed Firms*, Theoretical Economics Letters 8 (2018), 755-772, dostupné na internete (27.4.2025): https://www.researchgate.net/publication/323802778_Predicting_Accuracy_of_Valuation_Multiples_Using_Value_Drivers_Evidence_from_Indian_Listed_Firms

[29] Hamner, B., Frasco, M.: Metrics: Evaluation Metrics for Machine Learning, R manuál, 2018, dostupné na internete (4.5.2025): https://CRAN.R-project.org/package=Metrics

[30] Hanauer, M.X., Kononova, M., Rapp, M.S.: *Boosting agnostic fundamental analysis: Using machine learning to identify mispricing in European stock markets*, Finance Research Letters 48 (2022), dostupné na internete (27.4.2025): https://www.sciencedirect.com/science/article/pii/S1544612322001465#sec0008a

[31] Hoberg, G., Philips G.: *Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis*, The Review of Financial Studies 23 (2010), 3773-3881, dostupné na internete (6.5.2025): https://academic.oup.com/rfs/article-abstract/23/10/3773/1564619

[32] Chen, T. et al.: *xgboost: Extreme Gradient Boosting*, R manuál, 2024, dostupné na internete (4.5.2025): https://CRAN.R-project.org/package=xgboost

[33] Chen, T., Guestrin, C.: *XGBoost: A scalable tree boosting system*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016), 785 - 794, dostupné na internete (18.1.2025) : https://dmlc.cs.washington.edu/data/pdf/XGBoostArxiv.pdf

[34] Chih-Ling T., Xiaogang S., Xin Y.: *Linear regression*, WIREs Computational Statistics 4 (2012), dostupné na internete: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.1198

[35] Chikodili, N. B., Mohammed, A. D., Opeyemi, A.: *Outlier Detection in Multivariate Time Series Data Using a Fusion of K-Medoid, Standardized Euclidean Distance and Z-Score*, Information and Communication Technology and Applications, príspevok na konferenciu ICTA 2020, https://www.researchgate.net/publication/349298548_Outlier_Detection_in_Multivariate_Time_Series_Data_Using_a_Fusion_of_K-Medoid_Standardized_Euclidean_Distance_and_Z-Score

[36] Interactive Brokers: *stránka na výpočet ceny SHORT pozície*, dostupné na internete (8.5.2025): https://www.interactivebrokers.com/en/pricing/short-sale-cost.php

[37] Jeffrey, R. A., Ulrich, J. M.: *quantmod: Quantitative Financial Modelling Framework*, R manuál, 2024, dostupné na internete (4.5.2025): https://CRAN.R-project.org/package=quantmod

[38] Jolliffe, I. T., Cadima J.: *Principal component analysis: a review and recent developments*, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374 (2016),dostupné na internete (30.4.2025): https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202

[39] Kassambara, A.: *ggpubr: 'ggplot2' Based Publication Ready Plots*, R manuál, 2023, dostupné na internete (4.5.2025): https://CRAN.R-project.org/package=ggpubr

[40] Kassambara, A.: *ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'*, R manuál, 2023, dostupné na internete (4.5.2025): https://CRAN.R-project.org/package=ggcorrplot

[41] Koller, T. et al.: *Valuation: Measuring and Managing the Value of Companies*, John Wiley & Sons, Hoboken, 2015

[42] Krüger, P., Landier, A., Thesmar, D.: *The WACC Fallacy: The Real Effects of Using a Unique Discount Rate*, The Journal of Finance 70 (2015), 1253-1285, dostupné na internete: https://www.jstor.org/stable/43611096

[43] Kuhn, M: *Building Predictive Models in R Using the caret Package*, Journal of Statistical Software 28 (2008), 1-26, dostupné na internete (4.5.2025): https://www.jstatsoft.org/index.php/jss/article/view/v028i05

[44] Kumar, A.: TheStoryofXGBoost:AmachineLearningRevolution, vedecko-populárny článok na internete, dostupné na internete (29.4.2025): https://dev.to/aashwinkumar/the-story-of-xgboost-a-machine-learning-revolution-bib

[45] Lee, Ch.M.C., Ma,P., Wang, Ch.C.Y: *Search-based peer firms: Aggregating investor perceptions through internet co-searches*, Journal of Financial Economics 116 (2015), 410-431, dostupné na internete: https://www.sciencedirect.com/science/article/pii/S0304405X15000197

[46] Li, S., Zhang, X.: Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm, Neural Computing and Applications 32 (2020),1971 - 1979, dostupný na internete (29.4.2025): https://link.springer.com/article/10.1007/s00521-019-04378-4

[47] *Makkot3a*, preložené hebrejské náboženské texty, dostupné na internete (5.5.2025): https://www.sefaria.org/Makkot.3a.14?lang=bi&with=all&lang2=en

[48] Malpani, K: *Detecting Outliers for Single Dimensional Data Using Interquartile Range*, Journal of Engineering Research and Application 9 (2019), 31-35, dostupné na internete (6.5.2025): InterquartileRange

[49] Modigliani, F., Miller, M.H.: *The Cost of Capital, Corporation Finance and the Theory of Investment*, The American Review 48 (1958), 261-297, dostupné na internete (29.4.2025): https://www.jstor.org/stable/1809766

[50] Neeraj, J.: *Portfolio Optimization Using R (HM Model)*, vedecko-popularizačné články, dostupné na internete (5.5.2025): PortfolioOptimizationUsingR(HMModel)

[51] Nobre, J., Neves, R.F.: *Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to trade in the financial markets*, Expert Systems with Applications 125 (2019), 181-194, dostupné na internete (29.4.2025): https://www.sciencedirect.com/science/article/abs/pii/S0957417419300995

[52] Ooms, J: *writexl: Export Data Frames to Excel 'xlsx' Format*, R manuál, 2024, dostupné na internete: https://CRAN.R-project.org/package=writexl

[53] Petersen, T.L.: *patchwork: The Composer of Plots*, R manuál, 2024, dostupné na internete: https://CRAN.R-project.org/package=patchwork

[54] Rudin, C.: *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*, Nature Machine Intelligence 1 (2019), 206-215, dostupné na internete (7.5.2025): https://www.nature.com/articles/s42256-019-0048-x

[55] Sharpe, W.F.: *Capital Asset Prices: A theory of market equilibrium under conditions of risk*, The Journal of Finance 19 (1964), 425-442, dostupné na internete: https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.1964.tb02865.x

[56] Tay, F.E.H., Cao, L.J.: *Modified support vector machines in financial time series foecasting*, Neurocomputing 48 (2002), 847-861, dostupné na internete (29.4.2025): https://www.sciencedirect.com/science/article/abs/pii/S0925231201006762

[57] Üstün,B., Melssen, W.J., Buydens, L.M.C.: *Visualisation and interpretation of Support Vector Regression models* 595 (2007), 299-309, dostupné na internete (1.5.2025): https://www.sciencedirect.com/science/article/abs/pii/S0003267007004904

[58] *výnosy amerických dlhopisov*, dostupné na internete (1.5.2025): https://www.investing.com/rates-bonds/u.s.-10-year-bond-yield-historical-data

[59] Wickham, H: *ggplot2: Elegant Graphics for Data Analysis*, Springer, New York, 2016

[60] Wickham, H. et al.: *dplyr: A Grammar of Data Manipulation*, R manuál, 2023, dostupné na internete: https://CRAN.R-project.org/package=dplyr

[61] Wickham, H., Vaughan, D., Girlich, M.: *tidyr: Tidy Messy Data*, R manuál, 2024, dostupné na internete: https://CRAN.R-project.org/package=tidyr

[62] Wickham, H.: *forcats: Tools for Working with Categorical Variables (Factors)*, R manuál, 2023, dostupné na internete (4.5.2025) https://CRAN.R-project.org/package=forcats

[63] Wickham, H., Pedersen, T. L., Seidel, D.: *scales: Scale Functions for Visualization*, R manuál, 2023, dostupné na interne (4.5.2025): https://CRAN.R-project.org/package=scales

[64] Weingessel, A.: *Functions to Solve Quadratic Programming Problems*, R manuál, 2019, dostupné na internete (4.5.2025): https://CRAN.R-project.org/package=quadprog

[65] Zivot, E.: *Introduction to Computational Finance and Financial Econometrics with R*, online učebnica, University of Washington, Washington, 2016, dostupné na internete (8.5.2025): https://bookdown.org/compfinezbook/introcompfinr/

Príloha A: Naivne diverzifikované portfólio

Pri zostavovaní kódov sme vychádzali zo zdrojov [5], [43], [32], [59], [60], [61], [52], [39], [40], [62], [2], [53], [37], [63], [27], [64], a [15].

```
################################################
# NAIVNE DIVERZ. PORTFOLIO
################################################
# install.packages(c("readxl","quantmod"))
# library(readxl); library(quantmod)

# 1) Pracovn  adres r a tickery
setwd("/Users/michaldrgon/Desktop/Diplomovka/K dy")
data    <- read_excel("TICKERS.xlsx", col_names = TRUE)
tickers <- data[[1]]
# 2) Na e dva d tumy
dates <- as.Date(c("2023-12-31","2024-12-31"))
################################################
#funckia na s ahovanie
# niekolko tickerov robilo problemy
#dal som im ignor, aby sa stiahol zvysok
################################################
# samotna funkcia na stahovanie
get_one_adj_after <- function(sym, D, lookahead = 10) {
  tmp_env <- new.env()
   ok <- tryCatch({
suppressWarnings(
getSymbols(sym,
src         = "yahoo",
from        = D,
to          = D + lookahead,
auto.assign = TRUE,
env         = tmp_env,
warnings    = FALSE,
verbose     = FALSE)
```

```
)
TRUE
}, error = function(e) FALSE)

  if (!ok) return(NA_real_)

  # iba adjusted stlpec
  adj_xts <- tryCatch(Ad(tmp_env[[sym]]), error = function(e)
     NULL)
  if (is.null(adj_xts)) return(NA_real_)

 # vyberie prvu hodnotu ktora je dostupna do D, alebo vratane
idx <- index(adj_xts) >= D
if (!any(idx)) return(NA_real_)

  as.numeric(adj_xts[idx][1])
}

# 4) prirpava datoveho ramca na spracovanie
result <- expand.grid(Ticker = tickers, Date = dates,
                      stringsAsFactors = FALSE)
result$Adjusted <- NA_real_

for (i in seq_len(nrow(result))) {
  sym <- result$Ticker[i]
  D   <- result$Date[i]
  result$Adjusted[i] <- get_one_adj_after(sym, D, lookahead = 10)
}
print(result)

#library(dplyr)
#library(tidyr)
###################################
```

```r
# vdaka tymto knizniciam pojde lahsie
# maniupolvat s datami
###################################

# 6) Pivot do wide, st pce bud  napr. "Price_2023-12-31" a "Price_2024-12-31"
result_wide <- result %>%
  pivot_wider(names_from  = Date,
              values_from = Adjusted,
              names_prefix = "Price_")

# 7) lepsie pomenovanie, aby som sa vyznal na "Price_2023" a "Price_2024"
dates_chr <- format(dates, "%Y-%m-%d")
old_names <- paste0("Price_", dates_chr)
new_names <- c("Price_2023", "Price_2024")
names(result_wide)[ names(result_wide) %in% old_names ] <- new_names
# 8) vypocet jednoducheho returnu
result_wide <- result_wide %>%
  mutate(Return = (Price_2024 / Price_2023) - 1)
naive_return <- mean(result_wide$Return, na.rm = TRUE)
# 9) V pis
print(result_wide)
cat("V nos naivn ho portf lia:", round(naive_return * 100, 2), "%\n")
```

Príloha B: Markowitzovo portfólio

```
################################################################
# MARKOWITZ - GMV + MV (S)
################################################################
cesta_suboru          <-
   "/Users/michaldrgon/Desktop/Diplomovka/K dy/TICKERS.xlsx"
testovaci_rezim       <- FALSE
pocet_testovacich     <- 3
zaciatok_odhadov      <- "2023-01-01"
koniec_odhadov        <- "2023-12-31"
datum_nakupu_min      <- "2023-12-31"
datum_predaja_min     <- "2024-12-31"
pocet_obchodnych_dni  <- 252
povolit_shortovanie   <- TRUE
cov_reg               <- 1e-6

################################################################
# Bal   ky
################################################################
library("readxl")
library("dplyr")
library("tidyr")
library("tidyquant")
library("quadprog")
library("ggplot2")
library("scales")

################################################################
# Na   tanie  d  t  a  v  po  et  mu a Sigma
################################################################
tickery <- readxl::read_excel(cesta_suboru,
   col_names=FALSE)[[1]] %>%
  na.omit()
```

```
ceny <- tidyquant::tq_get(
  tickery, get="stock.prices",
  from=zaciatok_odhadov, to=koniec_odhadov
) %>%
  select(symbol, date, adjusted) %>%
  arrange(symbol, date)
navraty <- ceny %>%
  group_by(symbol) %>%
  mutate(simple_ret = adjusted/lag(adjusted) - 1) %>%
  ungroup() %>%
  select(date, symbol, simple_ret) %>%
  pivot_wider(names_from=symbol, values_from=simple_ret) %>%
  drop_na()
matica_ret <- navraty %>% select(-date) %>% as.matrix()
symboly <- colnames(matica_ret)
n <- length(symboly)
mu_daily    <- colMeans(matica_ret)
Sigma_daily <- cov(matica_ret) + diag(rep(cov_reg, n))
mu_ann    <- mu_daily    * pocet_obchodnych_dni
## kvoli tomu, ze je tych firiem vela, tak kvoli numerickej
   stabilite sa pridava mala
# konstantna aby neboli problemy s invert
Sigma_ann <- Sigma_daily * pocet_obchodnych_dni + cov_reg *
   diag(n)

################################################################
# GMV portf lio
################################################################
Dmat_g <- 2 * Sigma_ann
dvec_g <- rep(0, n)
if(povolit_shortovanie){
  Amat_g <- matrix(1, nrow=n)
  bvec_g <- 1
```

```
  meq_g  <- 1
} else {
  Amat_g <- cbind(rep(1,n), diag(n))
  bvec_g <- c(1, rep(0,n))
  meq_g  <- 1
}
w_gmv <- solve.QP(Dmat=Dmat_g, dvec=dvec_g, Amat=Amat_g,
   bvec=bvec_g, meq=meq_g)$solution
w_gmv <- w_gmv / sum(w_gmv)

gmv_weights <- tibble(symbol=symboly, vaha=w_gmv)
cat("GMV␣v hy:\n");
print(gmv_weights)

#################################################################
# S portfolio, CML
#################################################################
rf <- 0.03898  # yield US 10Y k 31.12.2023

w_raw     <- solve(Sigma_ann) %*% (mu_ann - rf)
w_tang    <- as.numeric(w_raw / sum(w_raw))
w_tang    <- w_tang / sum(w_tang)

tang_weights <- tibble(symbol=symboly, vaha=w_tang)
cat("S␣v hy:\n"); print(tang_weights)

#################################################################
# ocakavan  v nos
#################################################################
exp_ret_gmv  <- sum(w_gmv  * mu_ann)
exp_ret_tang <- sum(w_tang * mu_ann)
cat("O ak van ␣ro n ␣v nos␣GMV:␣␣␣",
   scales::percent(exp_ret_gmv),  "\n")
```

```r
cat("Očakávaný ročný výnos:",
    scales::percent(exp_ret_tang), "\n")

#################################################################
# realny expost vynos
#################################################################
ceny_nakupu <- tidyquant::tq_get(
  tickery, get="stock.prices",
  from=datum_nakupu_min, to=as.Date(datum_nakupu_min)+7
) %>% group_by(symbol) %>% slice_min(date) %>% ungroup() %>%
    rename(price_buy=adjusted)

ceny_predaja <- tidyquant::tq_get(
  tickery, get="stock.prices",
  from=datum_predaja_min, to=as.Date(datum_predaja_min)+7
) %>% group_by(symbol) %>% slice_min(date) %>% ungroup() %>%
    rename(price_sell=adjusted)


calc_expost <- function(weights){
  tibble(symbol=symboly, vaha=weights) %>%
    left_join(ceny_nakupu,  by="symbol") %>%
    left_join(ceny_predaja, by="symbol") %>%
    filter(!is.na(price_buy), !is.na(price_sell), price_buy>0)
       %>%
    mutate(ret_i   = price_sell/price_buy - 1,
           ret_pos = vaha * ret_i) %>%
    summarise(expost = sum(ret_pos, na.rm=TRUE)) %>%
    pull(expost)
}

real_ret_gmv  <- calc_expost(w_gmv)
real_ret_tang <- calc_expost(w_tang)
```

```r

cat("expost␣vynos␣GMV:␣␣␣", scales::percent(real_ret_gmv),  "\n")
cat("expost␣vynos␣S:", scales::percent(real_ret_tang), "\n")

############################################################
# expozicia hruba aj cista vypocet
############################################################
net_gmv   <- sum(w_gmv)
gross_gmv <- sum(abs(w_gmv))
net_tang   <- sum(w_tang)
gross_tang <- sum(abs(w_tang))

cat("GMV␣Net␣EXP:␣␣␣", round(net_gmv, 3),   "\n")
cat("GMV␣Gross␣EXP:␣", round(gross_gmv, 3), "\n")
cat("S␣Net␣EXP:␣␣␣␣", round(net_tang, 3),   "\n")
cat("S␣Gross␣EXP:␣␣", round(gross_tang, 3), "\n")


############################################################
# vykreslenie EF + portfolia a CML
############################################################

risk_tang <- sqrt(as.numeric(w_tang %*% Sigma_ann %*% w_tang))
ret_tang  <- exp_ret_tang

##predlzenie osi
mu_min <- min(mu_ann)*0.8
mu_max <- max(mu_ann, ret_tang) * 1.1
mu_seq   <- seq(mu_min, mu_max, length.out = 2000)

#vykreslenie hranicee
one      <- rep(1, n); iS <- solve(Sigma_ann)
A        <- as.numeric(t(one) %*% iS %*% one)
```

```
B         <- as.numeric(t(one) %*% iS %*% mu_ann)
C         <- as.numeric(t(mu_ann) %*% iS %*% mu_ann)
D         <- A*C - B^2
risk_seq <- sqrt((A*mu_seq^2 - 2*B*mu_seq + C) / D)
frontier <- tibble(ret = mu_seq, risk = risk_seq,
                   branch = ifelse(mu_seq < B/A, "lower",
                      "upper"))

#nastavenie limit
x_lim <- c(0, risk_tang * 1.2)
y_lim <- c(rf * 0.8, ret_tang  * 1.2)

# 5) nakreslime
ggplot() +
  geom_line(data = filter(frontier, branch=="lower"),
            aes(risk, ret), color="grey60", size=0.5) +
  geom_line(data = filter(frontier, branch=="upper"),
            aes(risk, ret), color="steelblue", size=1) +
  # CML
  geom_abline(intercept = rf, slope = (ret_tang - rf)/risk_tang,
              linetype="dashed", color="darkgreen") +
  # GMV a S body
  geom_point(aes(x=sqrt(as.numeric(w_gmv %*% Sigma_ann %*%
     w_gmv)),
                 y=exp_ret_gmv),
             color="red", size=3) +
  geom_text(aes(x=sqrt(as.numeric(w_gmv %*% Sigma_ann %*%
     w_gmv)),
                y=exp_ret_gmv, label="GMV"),
            hjust=1.1, vjust=-0.5, color="red") +
  geom_point(aes(x=risk_tang, y=ret_tang),
             color="darkgreen", size=3) +
  geom_text(aes(x=risk_tang, y=ret_tang, label="S"),
```

```
185             hjust=-0.1, vjust=1.1, color="darkgreen") +
186   scale_x_continuous("Riziko␣  ␣(annualizovan )",
187                      labels = percent_format(0.1),
188                      limits = x_lim) +
189   scale_y_continuous("Ro  n  ␣jednoduch ␣v  nos",
190                      labels = percent_format(1),
191                      limits = y_lim) +
192   ggtitle("") +
193   theme_minimal()
```

Príloha C: Automatizovaný model a LONG-SHORT portfólio

```
##############################
# instalacia kniznic
##############################
# install.packages(c(
#   "readxl","caret","xgboost","ggplot2","Metrics",
#   "dplyr","tidyr","zoo","writexl","ggpubr","ggcorrplot","pdp"
# ))
# install.packages("pdp")
#install.packages("viridis")
library(viridis)
library(readxl)
library(caret)
library(xgboost)
library(ggplot2)
library(Metrics)
library(dplyr)
library(tidyr)
library(zoo)
library(writexl)
library(ggpubr)
library(ggcorrplot)
library(pdp)
library(patchwork)
library(data.table)

##############################
# funckie na odstranovanie odlah. pozorov.
##############################
remove_outliers_iqr <- function(df, cols) {
for (col in cols) {
```

```
Q1 <- quantile(df[[col]], .25, na.rm=TRUE)
Q3 <- quantile(df[[col]], .75, na.rm=TRUE)
IQR_val <- Q3 - Q1
df <- df %>% filter(.data[[col]] >= Q1 - 1.5 *
   IQR_val,.data[[col]] <= Q3 + 1.5 * IQR_val)
  }
  df
}
remove_outliers_pca <- function(df, cols, pca_components=NULL,
   threshold=.9) {
pca_res <- prcomp(df[, cols], scale.=TRUE)
scores  <- if (!is.null(pca_components)) pca_res$x[,
   1:pca_components] else pca_res$x
dist    <- sqrt(rowSums(scores^2))
cut_val <- quantile(dist, threshold, na.rm=TRUE)
df[dist <= cut_val, ]
}
##############################
# predspracovanie dat
##############################
data <-
   read_excel("/Users/michaldrgon/Desktop/Diplomovka/DATA_DIPL.xlsx",
   sheet="LOAD")
data$date <- as.Date(data$date_num, format="%m/%d/%y")
data$date_num <- NULL
colnames(data) <- make.names(colnames(data))

# casova premenna na [0,1]
data$day_count <- as.numeric(data$date - min(data$date))
data$day_count <- (data$day_count - min(data$day_count))
   /(max(data$day_count) - min(data$day_count))

##############################
```

```r
# 4. cielove premenne a features
##############################
targets <- c("market_to_book", "ev2a", "ev2s")
features <- setdiff(
names(data),c(targets, "date", "ticker", "Adjusted_Price",
"Gross.Profit.Margin", "Enterprise.value", "Free.Cash.Flow")
)
# Log-transform cie ov , filtrovanie NA/Inf a pridanie roka
data <- data %>%
mutate(across(all_of(targets), log)) %>%
filter(if_all(all_of(c(targets, features)), ~ is.finite(.) &
   !is.na(.))) %>%
mutate(year = format(date, "%Y"))
cat("Po et␣riadkov␣p vodne:", nrow(data), "\n")

##############################
# 4b. Defin cia mapy n zvov pre premenn
##############################
# Tu si dopl  v etky feature pod a svojho label_map
label_map <- c(
  "EBITDA.ratio"          = "EBITDA␣mar a",
  "Asset.Turnover"        = "␣ATR",
  "D2E"                   = "Pomer␣dlhu␣k␣vlastn mu␣imaniu",
  "Cash.Flow.Margin"      = "Mar a␣pe a n ch␣tokov",
  "Net.Profit.Margin"     = "  ist  ␣ziskov ␣mar a",
  "Operating.Profit.Margin" = "Prev dzkov ␣mar a",
  "ROA"                   = "ROA",
  "ROE"                   = "ROE",
  "day_count"            = "day_count"

)
##############################
## filterovanie outlierov
```

```
90  ##############################
91  data_iqr <- remove_outliers_iqr(data, c(targets,
        setdiff(features, "day_count")))
92  cat("Po et␣po␣IQR:", nrow(data_iqr), "\n")
93  data_pca <- remove_outliers_pca(data, c(targets,
        setdiff(features, "day_count")))
94  cat("Po et␣po␣PCA:", nrow(data_pca), "\n")
95
96  ##############################
97  # 6. korela n  anal za
98  ##############################
99  cor_mat_iqr <- cor(
100 data_iqr %>% select(all_of(features)) %>% select_if(is.numeric),
101 use = "pairwise.complete.obs")
102 ggcorrplot(cor_mat_iqr,lab = TRUE,method = "circle",
103 type = "lower",show.diag = FALSE,
104 colors = c("red", "white", "blue"),
105 title = "Korela n ␣matica␣|␣IQR")
106 cor_mat_pca <- cor(data_pca %>% select(all_of(features)) %>%
        select_if(is.numeric),
107 use = "pairwise.complete.obs")
108 ggcorrplot(cor_mat_pca,
109 lab = TRUE, method = "circle",
110 type = "lower",
111 show.diag = FALSE,
112 colors = c("red", "white", "blue"),
113 title = "Korela n ␣matica␣|␣PCA")
114
115 ##############################
116 #  vyhodnotenie
117 ##############################
118 tau <- 34.5
119 evaluate_model <- function(real, pred) {
```

```
r <- real
p <- pred
SSE  <- sum((r - p)^2, na.rm = TRUE)
SST  <- sum((r - mean(r, na.rm=TRUE))^2, na.rm = TRUE)
R2   <- 1 - SSE/SST
MedAE <- median(abs(r - p), na.rm = TRUE)
MedRE <- median(abs((r - p)/r), na.rm = TRUE) * 100
RMSE <- sqrt(mean((r - p)^2, na.rm = TRUE))
data.frame(R2 = R2,MedAE = MedAE,
MedRE = MedRE,
RMSE = RMSE)
}
##############################
run_modeling <- function(ds, filt) {
tr <- ds %>% filter(year <= "2022") %>% drop_na()
te <- ds %>% filter(year == "2023") %>% drop_na()
fu <- ds %>% filter(year == "2024") %>% drop_na()
summary_tbl   <- data.frame()
long_short    <- data.frame()
naive_tbl     <- data.frame()
inv_all       <- data.frame()
coef_tbl      <- data.frame()
importance_df <- data.frame()
ctrl <- trainControl(method = "cv", number = 5)
for (tg in targets) {
# -- LINE RNY MODEL --
lm_tr <- tr; lm_te <- te
for (col in features) {
m <- mean(tr[[col]], na.rm=TRUE)
lm_tr[[col]] <- tr[[col]] - m
lm_te[[col]] <- te[[col]] - m
}
```

```r
lm_mod <- lm(as.formula(paste(tg, "~", paste(features, collapse
   = "+"))), data = lm_tr)
lm_pr  <- predict(lm_mod, newdata = lm_te)
# ulozeenie koeficientov a vsetkych tabuliek
coef_df <- data.frame(
Target      = tg,
Filtration  = filt,
Variable    = names(coef(lm_mod)),
Coefficient = coef(lm_mod),
row.names   = NULL)
coef_tbl <- bind_rows(coef_tbl, coef_df)
summary_tbl <- rbind(
summary_tbl,
data.frame(Target = tg, Filtration = filt, Model = "Linear",
evaluate_model(te[[tg]], lm_pr))
# XGB
set.seed(123)
xgb_mod <- train(
x=as.matrix(tr[, features]),y= tr[[tg]], method    = "xgbTree",
trControl = ctrl,
tuneGrid  = expand.grid(
nrounds=c(100,300),
max_depth= c(3,5),
eta=c(0.03,0.1),
gamma=0,
colsample_bytree = 0.8,
min_child_weight = 1,
subsample= 0.8
),
verbosity = 0
)
# var importance
imp_mat <- xgb.importance(
```

```
feature_names = features,model= xgb_mod$finalModel)
imp_mat$Target      <- tg
imp_mat$Filtration <- filt
importance_df<- bind_rows(importance_df, imp_mat)
# 3) predikce a vyhodnocen
xgb_pr <- predict(xgb_mod, newdata = as.matrix(te[, features]))
summary_tbl <- rbind(
summary_tbl,
data.frame(
Target= tg,
Filtration = filt,
Model= "XGBoost",
evaluate_model(te[[tg]], xgb_pr)))
# -- SVM MODEL --
set.seed(123)
svm_mod <- train(
x= as.matrix(tr[, features]),
y= tr[[tg]],
method= "svmRadial",
preProcess = c("center","scale"),
trControl  = ctrl,
tuneLength = 3)
svm_pr <- predict(svm_mod, newdata = as.matrix(te[, features]))
summary_tbl <- rbind(summary_tbl,data.frame(Target = tg,
   Filtration = filt, Model = "SVM",
evaluate_model(te[[tg]], svm_pr)))
# LONG SHORT
for (method in c("XGBoost","SVM")) {
pr_col <- if (method=="XGBoost") xgb_pr else svm_pr
te$PR <- pr_col
p23 <- te %>% group_by(ticker) %>% filter(date==max(date)) %>%
   ungroup() %>% select(ticker, P23 = Adjusted_Price)
```

```
p24 <- fu %>% group_by(ticker) %>% filter(date==max(date)) %>%
   ungroup() %>% select(ticker, P24 = Adjusted_Price)
inv <- te %>% select(ticker, Pred = PR, Val = all_of(tg)) %>%
left_join(p23) %>% left_join(p24) %>%
filter(!is.na(P23) & !is.na(P24)) %>%
mutate(T_orig = exp(Val),
P_orig = exp(Pred),
Mis= (P_orig - T_orig)/T_orig*100,
Pos= case_when(Mis >= tau  ~ "LONG", Mis <= -tau ~ "SHORT",
TRUE~ "NEUTRAL"),
Conf    = pmax(0, abs(Mis) - tau))
inv <- if (sum(inv$Conf)>0) inv %>% mutate(W = Conf/sum(Conf))
   else inv %>% mutate(W = 1/n())
inv <- inv %>% mutate(Return = case_when(
Pos=="LONG"  ~ (P24/P23)-1,
Pos=="SHORT" ~ -(P24/P23)+1,
TRUE           ~ 0))
# overall return
long_short <- rbind(long_short, data.frame(Target = tg,
   Filtration = filt, Model = method,
Return = sum(inv$W * inv$Return)))
naive_tbl <- rbind(naive_tbl, data.frame(Target = tg, Filtration
   = filt, Model = method,
Naive_Return = mean((inv$P24/inv$P23)-1)))
inv$Model <- method; inv$Filtration <- filt; inv$Target <- tg
inv_all <- bind_rows(inv_all, inv)}}
return(list(
summary = summary_tbl,
long_short= long_short,
naive= naive_tbl,
investments  = inv_all,
coefficients = coef_tbl,
importance   = importance_df
```

```
  ))
}
##############################
# spustenie modelu
##############################
results_iqr <- run_modeling(data_iqr, "IQR")
results_pca <- run_modeling(data_pca, "PCA")

##############################
# v sledky
##############################
summary_table_all   <- bind_rows(results_iqr$summary,
   results_pca$summary)
long_short_all      <- bind_rows(results_iqr$long_short,
   results_pca$long_short)
naive_all           <- bind_rows(results_iqr$naive,
   results_pca$naive)
investments_all     <- bind_rows(results_iqr$investments,
   results_pca$investments)
coefficients_all    <- bind_rows(results_iqr$coefficients,
   results_pca$coefficients)

coefficients_all_trans <- coefficients_all %>%
mutate(Effect_dec = exp(Coefficient) - 1)

# --- NOV : v pis so zaokr hlen m ---
cat("\n*** Koeficienty line rnej regresie exp***\n")
print(coefficients_all_trans %>% mutate(Coefficient =
   round(Coefficient, 2),
Effect_dec  = round(Effect_dec, 2)))
##############################
# zlucenie
##############################
```

```
importance_all <-
   bind_rows(results_iqr$importance,results_pca$importance)
##############################
#  vyp sanie
##############################
cat("\n*** S hrn metr k ***\n");       print(summary_table_all)
cat("\n*** V nosy LONG-SHORT (celkovo) ***\n");
   print(long_short_all)
cat("\n*** Naivn  v nosy ***\n");      print(naive_all)
cat("\n*** Koeficienty line rnej regresie ***\n");
   print(coefficients_all)
##############################
# 13.rozklad long vs short
##############################
contrib_all <- investments_all %>%
group_by(Filtration, Model, Target, Pos) %>%
summarise(Weighted_Return = sum(W*Return, na.rm=TRUE), .groups =
   'drop') %>%
pivot_wider(names_from = Pos, values_from = Weighted_Return,
   values_fill = 0)
cat("\n*** Rozklad v nosu LONG vs SHORT ***\n");
   print(contrib_all)
##############################
# 14.  sum r
##############################
contrib <- investments_all %>%
group_by(Filtration,Model,Target) %>%
summarise(Long_Return  = sum(ifelse(Pos=="LONG",  W*Return,0),
   na.rm=TRUE),
Short_Return = sum(ifelse(Pos=="SHORT", W*Return,0), na.rm=TRUE))
print(contrib)
manual_grid_pdp <- function(filtracia, target) {
ds <- if (filtracia == "IQR") data_iqr else data_pca
```

```
tr <- ds %>% filter(year <= "2022") %>% drop_na()
mod <- train(x= as.matrix(tr[, names(label_map)]),
y= tr[[target]],
method = "xgbTree",
trControl = trainControl(method = "cv", number = 5),
tuneGrid  = expand.grid(nrounds= 100,
max_depth        = 3,
eta              = 0.1,
gamma            = 0,
colsample_bytree = 0.8,
min_child_weight = 1,
subsample        = 0.8),verbosity = 0
  )
# 4) vypo  taj PDP pre ka d  premenn
pdp_list <- lapply(names(label_map), function(feat) {
pd <- pdp::partial(
object= mod$finalModel,
pred.var= feat,
train= tr[, names(label_map)],
grid.resolution = 300,
progress        = "none")
ggplot(pd, aes_string(x = feat, y = "yhat")) +
geom_line(size = 1) +
labs(x = label_map[[feat]], y = "Predikovan hodnota") +
theme_minimal() +theme(axis.title = element_text(size = 9),
axis.text  = element_text(size = 8))})
# vykreslenie
plots <- wrap_plots(pdp_list, ncol = 3)
  print(plots)
}
# vygenerovanie obrazkov
manual_grid_pdp("IQR", "market_to_book")
manual_grid_pdp("IQR", "ev2a")
```

```
manual_grid_pdp("IQR", "ev2s")
manual_grid_pdp("PCA", "market_to_book")
manual_grid_pdp("PCA", "ev2a")
manual_grid_pdp("PCA", "ev2s")



plot_importance_unified <- function(filtration, target, metric,
    fill_color) {
  df <- if (metric == "Gain") {
importance_all %>%
filter(Filtration == filtration, Target == target) %>%
mutate(Feature = fct_reorder(Feature, Gain), Value = Gain)
  } else {imp_long %>%
filter(Filtration == filtration, Target == target, Metric ==
    metric) %>%
mutate(Feature = fct_reorder(Feature, Value))
}
  if (nrow(df) == 0) {
message(" iadne ␣d ta␣pre␣kombin ciu:␣", filtration, ",␣",
    target, ",␣", metric)
     return(NULL)
  }

ggplot(df, aes(x = Feature, y = Value)) +
geom_col(fill = fill_color) +
coord_flip() +
scale_x_discrete(labels = label_map) +
labs(title = paste(""),
x = NULL, y = metric
) +
theme_minimal() +
```

```
359 theme(axis.text.y = element_text(size = 12))
360 }
361
362 importance_all <- bind_rows(results_iqr$importance,
      results_pca$importance)
363
364 imp_long <- importance_all %>%
365   pivot_longer(cols = c(Cover, Frequency), names_to = "Metric",
        values_to = "Value")
366
367
368
369 plot_importance_unified("IQR", "market_to_book", "Gain",
      "steelblue")
370 plot_importance_unified("IQR", "market_to_book", "Cover",
      "steelblue")
371 plot_importance_unified("IQR", "market_to_book", "Frequency",
      "steelblue")
372
373 plot_importance_unified("IQR", "ev2a", "Gain",      "steelblue")
374 plot_importance_unified("IQR", "ev2a", "Cover",     "steelblue")
375 plot_importance_unified("IQR", "ev2a", "Frequency", "steelblue")
376
377 plot_importance_unified("IQR", "ev2s", "Gain",      "steelblue")
378 plot_importance_unified("IQR", "ev2s", "Cover",     "steelblue")
379 plot_importance_unified("IQR", "ev2s", "Frequency", "steelblue")
380
381 plot_importance_unified("PCA", "market_to_book", "Gain",
      "tomato")
382 plot_importance_unified("PCA", "market_to_book", "Cover",
      "tomato")
383 plot_importance_unified("PCA", "market_to_book", "Frequency",
      "tomato")
```

```
384
385 plot_importance_unified("PCA", "ev2a", "Gain",      "tomato")
386 plot_importance_unified("PCA", "ev2a", "Cover",     "tomato")
387 plot_importance_unified("PCA", "ev2a", "Frequency", "tomato")
388
389 plot_importance_unified("PCA", "ev2s", "Gain",      "tomato")
390 plot_importance_unified("PCA", "ev2s", "Cover",     "tomato")
391 plot_importance_unified("PCA", "ev2s", "Frequency", "tomato")
392
393 allocation_all <- investments_all %>%
394 filter(Pos %in% c("LONG", "SHORT", "NEUTRAL")) %>%
395 group_by(Filtration, Model, Target, Pos) %>%
396 summarise(Total_Allocation = sum(W, na.rm = TRUE), .groups =
       "drop") %>%
397 pivot_wider(names_from = Pos, values_from = Total_Allocation,
       values_fill = list(Total_Allocation = 0)) %>%
398 mutate(
399 Total       = LONG + SHORT + NEUTRAL,
400 Long_pct    = round(100 * LONG / Total, 1),
401 Short_pct   = round(100 * SHORT / Total, 1),
402 Neutral_pct = round(100 * NEUTRAL / Total, 1)
403 ) %>%  select(Filtration, Model, Target, Long_pct, Short_pct,
       Neutral_pct)
404 # Vyp    v sledok
405 print(allocation_all)
406
407 # n vrotnosti, keby sme investovali iba do long
408 standalone_long <- investments_all %>%
409 filter(Pos == "LONG") %>%
410 group_by(Filtration, Model, Target) %>%
411 summarise(Standalone_Long = sum((W / sum(W)) * Return),.groups =
       "drop")
412
```

```
413 #n vrotnosti , keby sme investovali iba do short
414 standalone_short <- investments_all %>%
415 filter(Pos == "SHORT") %>%
416 group_by(Filtration, Model, Target) %>%
417 summarise(Standalone_Short = sum((W / sum(W)) * Return),.groups
      = "drop")
418
419 # tabulka vysledkov
420 standalone <- full_join(standalone_long, standalone_short,
421  by = c("Filtration","Model","Target"))
422 print(standalone)
```