Slovenska technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-100241-74044

Bc. Peter Križan

PREDIKCIA VÝVOJA CENOVEJ HLADINY BITCOINU

Priebežná správa o riešení DP2

Študijný program: Inteligentné softvérové systémy
Študijný odbor: Softvérové inžinierstvo – hlavný študijný odbor
Umelá inteligencia – vedľajší študijný odbor
Miesto vypracovania: Ústav informatiky, informačných systémov a softvérového Inžinierstva, FIIT STU v Bratislave
Vedúci práce: Ing. Jaroslav Loebl
December, 2019

Čestne vyhlasujem, že som túto prácu vypracoval samostatne, na základe konzultácií a s použitím uvedenej literatúry.

V Bratislave, 10.12.2019

Peter Križan

Anotácia

Slovenská technická univerzita v Bratislave

FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ

Študijný program: Inteligentné softvérové systémy

Autor: Bc. Peter Križan

Diplomová práca: Predikcia vývoja cenovej hladiny Bitcoinu

Vedúci diplomovej práce: Ing. Jaroslav Loebl

December, 2019

V tejto diplomovej práci sa venujeme téme predikcie vývoja cien kryptomeny Bitcoin. Samotné kryptomeny tvoria relatívne mladú položku na trhu s obchodovateľnými komoditami.. Ide o neregulovanú digitálnu menu, ktorej cenový vývoj zaznamenáva enormné výkyvy. Práve vďaka týmto výkyvom sa Bitcoin stáva lukratívnym prostriedkom trhovej výmeny. Princíp teórie efektívneho trhu (TET) definuje trh za efektívny a tým pádom bezpečným len v tom prípade, ak vývoj ceny nie je predikovateľný. V tejto práci využijeme doterajšie poznatky predchádzajúcich autorov a potvrdíme, prípadne vyvrátime TET. Nakoľko ide o digitálnu menu, zanecháva črty, ktoré boli podrobne preskúmané. Tieto črty majú rôzne formy vonkajšej prezentácie. Základnú sadu čŕt tvoria historické dáta, ktoré obsahujú doterajší vývoj ceny, ale taktiež aj záznamy uchovávané z technológie, ktorú Bitcoin využíva a to blockchain. Taktiež jednou z najdôležitejších časti práce je profilácia verejnej mienky na základe sentimentu extrahovaného zo sociálnej siete Twitter či Google Trends. Tieto dáta nám pomohli zistiť dopad kľúčových informácií na správanie sa verejnosti vo vzťahu k trhu a schopnosť šírenia sa týchto informácií. Navrhovaná metóda bude slúžiť k potvrdeniu, prípadne vyvráteniu TET v súvislosti s Bitcoinom.

Annotation

Slovak University of Technology Bratislava

FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

Degree course:	Intelligent Software Systems
Author:	Peter Križan
Master's thesis :	Prediction of Bitcoin price level
Supervisor:	Jaroslav Loebl

December, 2019

In this diploma thesis we deal with the topic of price level of bitcoin. The cryptocurrency are a relatively young item in the market for tradable commodities. This is an unregulated digital currency whose price fluctuates enormously. It is through these fluctuations that bitcoin becomes a lucrative means of market exchange. The principle of efficient market theory (TET) defines the market as efficient and thus safe only if the price development is not predictable. In this work we will use previous knowledge of authors and we will eventually corroborate or refute TET. As it is a digital form of money, it leaves features that we examine in detail. These features will have different forms of presentation. The basic set of features consists of historical data that contain the current development of the price, but also records kept from the technology used by bitcoin, namely blockchain. Also one of the most important parts of the work will be public opinion profiling based on the sentiment extracted from social networks or Google Trends. This data will help us to identify the impact of key information on public market behavior and the ability to disseminate this information. The proposed method will serve to confirm or refutation of TET in relation to bitcoin.

Obsah

1. Úvod

Už od nepamäti ľudstvo pociťovalo potrebu zavedenia všeobecne uznávanej komodity, ktorá by slúžila ako prostriedok všeobecnej výmeny. Prostriedok, ktorý by dokázal poskytovať hodnotu i napriek jej nepotrebnosti. V minulosti tieto prostriedky tvorili všeobecne využiteľné komodity ako bavlna, dobytok či kovy. Avšak i tieto prvopočiatočné platidlá sa časom ukázali ako nedostatočné, lebo mali slabú alebo takmer žiadnu trvácnosť. Od týchto čias sme sa pohli míľovými krokmi a už takmer 7000 rokov poznáme pojem peniaze [1].

V dnešnej dobe, slovo peniaze je veľmi abstraktný pojem. Okrem klasických papierových peňazí si asociujeme pod ním svoje bankové účty, rôzne druhy fondov, akcií alebo nehnuteľností. V každom z týchto prípadov však vkladáme peniaze do rúk autority alebo investujeme do relatívne hmotnej komodity s víziou jej budúceho nárastu hodnoty. V poslednom desaťročí však na trh vstúpila nová komodita - Bitcoin. Presne 3. januára 2009 bol vytvorený prvý blok v blockchaine a za jeho autora sa považuje Satoshi Nakamoto [2]. Toto meno však oficiálne neexistuje a doposiaľ nie je známy pravý autor či autori blockchainu pre Bitcoin. Krátko po uvedení Bitcoinu na trh sa zakladajú nové meny tohto druhu. Nakoľko ide o digitálne platidlá založené na kryptografií, globálne sú označované za kryptomeny.

K všeobecnému uznaniu niektorých kryptomien ako relevantného platidla prispieva najmä ich popularita. Uvedenie tohto typu platidla zaznamenalo veľký záujem u investorov a postupom času sa z nevýrazného digitálneho platidla stal objekt najčastejšie vyhľadávaných fráz v internetových vyhľadávačoch. Vplyv a značka technológie blockchain vyniesla hodnotu Bitcoinu z úvodných 0.06$ na rekordných 19 783.06$ dosiahnutých v roku 2017. V tomto období Bitcoin vzbudzuje záujem najširšieho okruhu záujemcov a stáva sa predmetom rôznych výskumov i všeobecnej kritiky.

Najväčšiu vlnu kritiky Bitcoin prijíma zo strany investičných bánk a ich zástupcov, nakoľko kryptomeny vtrhli na trh spôsobom, ktorý doteraz nezaznamenala žiadna iná komodita. Taktiež mnohé výskumy zaoberajúce sa touto tematikou naznačujú možnosť predpovedania vývoja cien tejto komodity. Tieto zistenia výrazne narúšajú samotnú podstatu trhu. Už v roku 1965 E. Famy uviedol svoju všeobecne uznanú prácu [3] o efektívnej teórii trhu, kde definuje trh ako subjekt, ktorý veľmi rýchlo spracováva nové informácie a nevykazuje rozdiel medzi akciovým kurzom a vnútornou hodnotou. V princípe definuje zmeny cien trhu

ako náhodné, kde príčinou zmeny kurzov je náhodná veličina. Z dôvodu, že tieto informácie nie sú vopred známe, vývoj ceny trhu musí byť čisto náhodný.

V závislosti od tohto tvrdenia sa otvára množstvo nepreskúmaných možností pre ciele ďalšieho výskumu. Kryptomeny, ako trhová komodita, bola spoločnosťou uznaná a denne s ňou obchodujú milióny ľudí. Preto i témou diplomovej práce bude oblasť kryptomien, proces vývoja a predikcie ich budúcich cien. Hlavný rozdiel oproti klasickým trhom vidíme v konečnom objeme obchodovateľnej komodity, nakoľko počet Bitcoinov bol vopred určený na 21 miliónov a taktiež ešte stále relatívne úzkym okruhom používateľov danej meny. Tieto fakty môžu mať za následok nie úplnú nezávislosť danej meny v trhovej ekonomike. Preto sa náš výskum pokúsi potvrdiť alebo vyvrátiť teóriu efektívneho trhu v tejto oblasti. S touto komoditou prichádza do styku iba určitá skupina ľudí a ich okruh je neporovnateľne menší ako pri iných trhoch. Preto sa zameriame na viaceré aspekty, ktoré môžu ovplyvňovať vývoj cien kryptomien. Či už pôjde o analýzu historických dát, mieru sentimentu na sociálnych sieťach, webových publikáciách alebo previazanosť tohto trhu s inými trhovými segmentami ako napríklad ropou. Analyzované poznatky využijeme pri snahe predikcie vývoja ceny kryptomien, ktorá v prípade preukázania kladnej bilancie vyvráti efektivitu tohto trhu.

2. Analýza

V tejto kapitole uvedieme čitateľa do problematiky, ktorú skúmame. Budú tu vymedzené pojmy, s ktorými sa čitateľ môže v našej práci stretnúť, ale aj analýza prác iných autorov zaoberajúcich sa rovnakou alebo obdobnou problematikou. Tieto práce budú poskytovať relevantný základ, ktorý bude nami preskúmaný, prípadne obohatený o nami doplnené poznatky. Taktiež sa pokúsime vždy asociovať zistené poznatky na náš problém, prípadne odvodiť použiteľné koncepty.

2.1. Úvod do pojmov

Nakoľko doména práce nie je čisto z oblasti informatiky, pokúsime sa vymedziť viacero pojmov, na ktorých znalosti bude táto práca postavená. Ide o pojmy z oblasti informatiky ale i z oblasti ekonómie, či trhovej ekonomiky nakoľko naša téma sa priamo dotýka oboch spomenutých oblastí.

Blockchain technológia bola prvýkrát predstavená autorom Satoshi Nakamoto v jeho publikácií [2], ktorá vytvorila základ kryptomeny známej ako Bitcoin. Táto publikácia nikdy nebola predložená tradičným recenzentským autoritám a autor túto publikáciu napísal pod pseudonymom. Autor v tejto práci nepredstavil len Bitcoin, ale aj technológiu, ktorá môže byť v širokom zmysle využitá vo finančnom sektore a tento princíp môže byť extrahovaný pre rôzne použitia. Hoci samotný autor nikde v svojej publikácií neuvádza pojem blockchain tento názov sa nesie v súvislosti s hlavnou ideou samotnej technológie, ktorá funguje na vzájomne prepojených blokoch a vytvárajú takzvanú reťaz blokov – blockchain. Autor prostredníctvom blockchain technológie vyriešil hlavný problém dôvery distribuovaných systémov. Hlavným princípom je, že žiadna zo zúčastnených strán nemôže manipulovať s obsahom údajov alebo s časovou pečiatkou bez detekcie. Principiálne si môžeme blok predstaviť ako transakciu obsahujúcu tri stĺpce, kde prvý stĺpec obsahuje časovú pečiatku danej transakcie, druhý stĺpec obsahuje samotnú transakciu a jej podrobnosti a tretí stĺpec obsahuje haš hodnotu predchádzajúceho bloku. Jediný spôsob akým by sa dalo manipulovať s údajmi bez detekcie, je nájsť kolíziu pri výpočte haš hodnôt, čo je prakticky nemožné [2].

Kryptomena, ako aj jej samotný názov naznačuje, je digitálna mena založená na kryptografickom princípe, ktorý zabezpečuje bezpečnú a ťažko sfalšovateľnú identitu. Azda najvýznamnejšou črtou kryptomien je ich decentralizácia. Nakoľko kryptomeny sú založené na blockchain prístupe vyznačujúci sa peer-to-peer prístupom nie sú regulované žiadnou centrálnou autoritou. Tento prístup chráni kryptomeny pred zásahmi tretích strán ako napríklad vlády, prípadne monopolov. Ďalšou z výhod blockchainových transakcií je prenos

zabezpečovaný verejnými a súkromnými kľúčmi, za cieľom absolútnej anonymity. Taktiež je možné prevody vykonať s minimálnymi poplatkami za spracovanie čo umožňuje vyhnúť sa poplatkom, ktoré sú účtované bežnými finančnými inštitúciami [18].

Trh je tá oblasť ekonomiky, kde dochádza k výmennej činnosti medzi jednotlivými subjektmi, prostredníctvom výmeny tovarov a služieb za trhové ceny [19]. V našom kontexte sa trh kryptomien najviac približuje devízovému trhu, preto i jednotlivé poznatky zistené pri snahe predikovať devízové trhy, budú hlbšie analyzované a bude posúdená ich vhodnosť aplikácie na trhy kryptomien.

Teória efektívneho trhu (TET) predpokladá, že kurzy sú ovplyvňované očakávanými ziskami, dividendami, rizikom a ďalšími kurzotvornými informáciami. Za efektívny je považovaný taký trh, ktorý veľmi rýchlo a presne absorbuje nové informácie. V situácii, keď všetky kurzotvorné informácie sú absorbované kurzom, tak nedochádza k rozdielu medzi vnútornou hodnotou a akciovým kurzom. Trhová cena na trhu predstavuje objektívnu hodnotu, komodity sú správne ocenené a na trhu nie je možné nájsť podhodnotené alebo nadhodnotené komodity. Termín efektívny sa teda používa v zmysle efektívneho spracovania nových informácií [3]. Efektívne chovanie akciových kurzov bolo už v minulosti skúmané. Obrovský význam pre TET mala práca [3], ktorej výsledkom je záver, že trhové kurzy sa chovajú náhodne. Táto práca sa stala zlomom, od ktorého sa datuje vznik TET.

V sfére kryptomien sú činitele ovplyvňujúce trh veľmi ťažko odhaliteľné, nakoľko ide o menu, ktorá je v dnešnej dobe využívaná celosvetovo za rôznymi účelmi. Častokrát býva platidlom v kriminálnej činnosti z dôvodu anonymity účastníkov danej transakcie.

Trhový sentiment podľa [4] vyjadruje globálne emócie účastníkov trhu v danom trhovom segmente. Okrem fundamentálnych a technických faktorov, tvorí mnohokrát opomínanú časť chovania sa daného trhu. Ako sme vyššie spomenuli efektívny trh je taký, ktorý dokáže rýchlo reagovať na jednotlivé podnety. V tomto prípade trhový sentiment zohráva výraznú rolu, nakoľko preklenuje dobu, kedy vzniká nový podnet z externých vplyvov a spoločnosť už reaguje na túto skutočnosť a výrazne tým ovplyvnia celkový dopad tohto podnetu. Reálnym príkladom môžu byť americké voľby prezidenta, kedy trhy prudko reagovali na víťazstvo Donalda Trumpa [4]. Nakoľko prieskumy odhadovali víťazstvo iného kandidáta, trhy ani účastníci trhov nereagovali na diametrálne odlišné politické názory tohto kandidáta. Po vyhlásení oficiálneho víťaza volieb, akcionári prudko reagovali na správu a práve šírenie tejto myšlienky, vyprofilovalo celkový dopad na trh. Trh bol však umelo ovplyvnený masovým počínaním si drvivej väčšiny akcionárov a umelo prehĺbili dopad tejto skutočnosti.

Netrvalo dlho aby sa trh opätovne stabilizoval, nakoľko reálne dopady neboli také výrazné ako ich médiá a spoločnosť vyprofilovali [4]. Tento príklad môže byť dôkazom toho, že nie len fundamentálne faktory dokážu výrazným spôsobom ovplyvniť trhovú situáciu ale aj sentiment verejnosti či schopnosť šírenia sa správ o danom trhovom segmente.

2.2. Príbuzné práce

Ako sme už vyššie spomenuli, viacero autorov sa pokúšalo predikovať ceny kryptomien. V analýze však začneme od prvopočiatkov vzniku idey predikovať cenu trhovej komodity. Od roku 1965, kedy bola položená TET, bolo viacero pokusov o vyvrátenie tejto teórie. Prvými, ktorým sa podarilo spochybniť TET sú Pruitt a White vo svojej práci [5]. Od tej doby sa prístupy k dolovaniu dát potrebných pre predikciu mien či iných typov komodít mnohonásobne rozrástli. Najnovšie štúdie naznačujú, že aj verejný názor obyvateľstva môže predstavovať výraznú informačnú stopu pre predikciu stavu cien trhových komodít. Jednotlivé práce autorov, by sa dali rozdeliť podľa obsahu, ktorý použili k predikovaniu cien kryptomien na trhu do nasledujúcich skupín:

- historické dáta,
- trendy vo vyhľadávaní,
- sentiment.

2.2.1. Historické dáta

Ako prvý a pravdepodobne najdôležitejší zdroj dát k úspešnej predikcii sú historické dáta. V roku 1988 autori Pruitt a Stephen W. publikovali článok [5], kde odhaľujú svoj systém pre predikciu finančných komodít s názvom CRISMA. Využívajú pri tom tri najbežnejšie filtre tej doby s použitím historických dát. Konkrétne ide o relatívnu silu, kumulatívny objem a 50 a 200 dňové kĺzavé priemery cien. Autor už v tej dobe necháva systém rozhodovať o nákupe a predaji komodít podľa spomenutých filtrov. V texte uvádza, že najdlhšia perióda držania komodity bola 113 dní a najkratšia iba 2 dni, čo dokazuje efektívnu kombináciu oboch obchodných stratégií a to kúp a drž (z angl. buy and hold) alebo využitie krátkodobých výkyvov na trhu. Celkovou kombináciou dokázal systém CRISMA vykazovať kladnú bilanciu a to aj pri započítaní 2% ceny za danú transakciu. Ako sme uviedli, prístup, ktorý zvolili bol čisto štatistický a v dnešnej dobe môže byť obohatený o výrazne silnejšie prediktory.

Zaujímavú myšlienku vyslovil Tom Bell v práci [6], kde naznačil previazanosť ceny Bitcoinu s cenami hardwérových súčastí potrebných k ťažbe. Ťažba v oblasti kryptomien

znamená poskytnutie hardvéru za účelom participácie na výpočte haš hodnôt, za ktoré je majiteľ hardvéru, baník, odmeňovaný. Autor vyslovil predpoklad, že vysoká cena Bitcoinu motivuje ku kúpe hardvéru potrebného k ťažbe a opačný efekt kedy vzrastajúca zložitosť ťažby má dopad na cenu, čím sa ťažba stáva pre väčšinu baníkov nezaujímavá. Autor tiež poskytol analýzu 30 – 100 dní kde preukázal vyššie spomenutú previazanosť. Tento prístup a použité dáta síce sú silne korelujúce, avšak narážame na nereálnosť získania takýchto dát v globálnej mierke. Zároveň uvádza dva najčastejšie spôsoby pri predikcií cien valút:

- **fundamentálna analýza** je technika, ktorá používa podkladové faktory zabezpečenia na odhad svojej hodnoty. Vo vzťahu k štátom emitovaným menám sa táto technika zameriava na metriky, ako sú prognózy rastu krajiny, úrovne dovozu a vývozu, cestovný ruch, politické opatrenia, úrovne dlhu, HDP a medzinárodné vzťahy. Používajú sa ako parametre modelu oceňovania. Ak je táto mena považovaná za podhodnotenú, potom má zmysel kupovať túto menu, inak predávať. V prípade kryptomien môžeme považovať fundamentálne črty historické dáta danej meny. Azda najznámejším fundamentálnym atribútom sú historické dáta, ktoré predstavujú zdrojový atribút a smerom k budúcnosti atribút cieľový [6].
- **Technická analýza** je alternatívnou metódou priradenia hodnoty zásobám, ktoré analyzujú činnosť trhu analýzou údajov, ako sú historické (kumulované) ceny a denný objem obchodov. Tento prístup sa nesnaží merať vnútornú hodnotu, ale používa matematické modely a štatistickú analýzu na identifikáciu modelov s cieľom predpovedať budúcu aktivitu [6].

Čisto historickými dátami (cenami) sa zaoberali aj autori Sean McNally et al. [7] kde dosiahli úspešnosť (angl. accuracy) klasifikácie 52%. Jednou zo základných čŕt je jednoduchý kĺzavý priemer (SMA) použitý pri metódach hlbokého učenia.

Jednoduchý kĺzavý priemer (SMA) – je aritmetický pohyblivý priemer. Je vypočítaný pridaním posledných (uzatváracích) cien a následným vydelením počtom časových období v sledovanom období. Použitím SMA sa dajú profilovať pohyby trhu v závislosti od sledovaného časového obdobia. Zatiaľ čo krátke časové obdobia sú vysoko citlivé na zmeny jednotlivých cien a dobre reagujú na zmeny posledných období, kým dlhšie sledované obdobia sú imúnne na krátkodobé výkyvy daného trhu [6]. Viacero autorov preto využíva kombináciu viacerých časových okien. Tento prístup považujeme za opodstatnený pri sledovaní vývoja cien na akomkoľvek trhu.

Autor využíva rekurentné neurónové siete s aplikovaným včasným zastavením na základe piatich po sebe nezlepšujúcich sa epochách. Štruktúra sledovaných modelov spočívala v

dvoch skrytých vrstvách s 50-100 neurónmi na každej. Taktiež preskúmal vplyv aktivačných funkcií na daný model. Ako funkciu s najlepším výsledkom autor označil tangens. Celkovo bol model testovaný na 10 – 10000 epochách, kde v druhom prípade autor hodnotí hrubé pretrénovanie modelu, ktoré autor na základe ďalších analýz vyriešil včasným zastavením trénovania ako sme už spomenuli vyššie. Okrem RNN modelu boli vyskúšané i ďalšie dva modely a to konkrétne LSTM a ARIMA. Výsledky všetkých modelov je možné vidieť v tabuľke 1 nižšie.

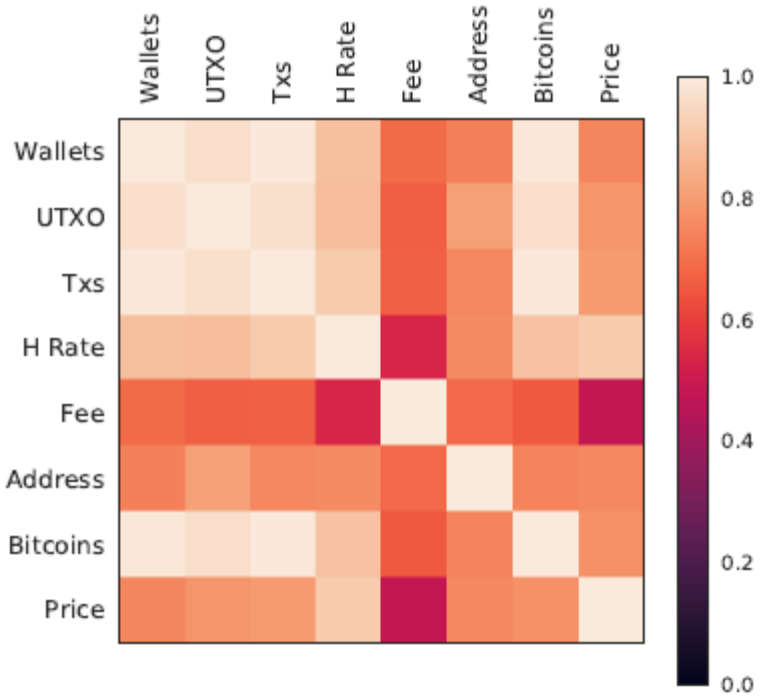
Tabuľka 1. Úspešnosť modelu dosiahnutá autormi [7] pre triedu nárast ceny

Model	**Časový úsek**	**Senzitivita**	**Presnosť** (z angl. precision)	**Úspešnosť** (z angl. accuracy)	**RSME**
LSTM	100	37%	35.50%	52.78%	6.87%
RNN	20	40.40%	56.65%	50.25%	5.45%
ARIMA	170	14.7%	100%	50.05%	53.74%

Historické dáta ponúkajú základnú škálu čŕt, ktoré môžeme sledovať, avšak ako aj autori prác [6][7] iba ojedinele využívali len tento druh dát. Väčšinou sú doplnené o dodatočné črty z iných segmentov. Zaujímavým faktom o historických dátach môžu byť taktiež informácie o blockchaine. Autori zväčša narábajú iba s hodnotou Bitcoinu ako s historickým atribútom. Online zdroj [29] ponúka kompletný vývoj blockchainu a obsahuje historické údaje o nasledujúcich atribútoch:

- hodnota odmien vyplatených baníkom,
- počet transakcií za deň,
- celková výstupná finančná hodnota za deň,
- odhadovaná hodnota transakcií v USD,
- celková hodnota odmeny za ťaženie vyplatená baníkom v USD,
- percentuálny podiel baníkov na dennom objeme transakcií,
- zložitosť počítaného hašu pre vytvorenie nového bloku,
- odhadovaný počet počítaných hašov za sekundu v sieti Bitcoin.

Jedným z autorov, ktorý zobral v úvahu historické údaje z blockchainu Bitcoinu je Muhammad Saad v práci [8], ktorý s využitím historických dát z blockchainu dokázal predikovať cenu Bitcoinu s úspešnosťou 99.4% . Autor previedol na začiatku svojej práce analýzu korelujúcich atribútov, ktorú je možné vidieť na obrázku1.



Obrázok 1. Korelačná matica pre atribúty blockchainu s trhovou cenou Bitcoinu [8]

Autor z vyššie spomenutej matice využíva iba atribúty dosahujúce korelačný koeficient vyšší ako 0.6. Celková úspešnosť 99.4% je však diskutabilná. Vzhľadom na veľmi krátky časový úsek sledovaných údajov, sme sa rozhodli analyzovať ceny Bitcoinu v autorom sledovanom období. Nižšie na obrázku 2 uvádzame graf cien Bitcoinu za autorom sledované obdobie. Ako je možné vidieť, graf má monotónne stúpajúcu tendenciu, čo vidíme ako jeden z dôvodov vysokej úspešnosti autora. Samotné sledované obdobie vykazuje extrémne málo údajov o poklese ceny. Zároveň autor neuvádza žiaden spôsob riešenia tejto situácie, v čom vidíme zásadný problém autorom prezentovaného riešenia. V spojení s najúspešnejším klasifikátorom, t. j. lineárnou regresiou, hodnotíme výsledný záver autora ako nie najlepší, nakoľko rozdelenie dátového priestoru do dvoch tried je vzhľadom na vstupné dané dáta triviálny problém. Našu teóriu podporujú aj závery zvyšných dvoch klasifikátorov a to konkrétne náhodný les, ktorý dosahuje úspešnosť 92.72%. Celkovo však významným zistením zostáva fakt, že aj historické dáta o samotnom blockchaine sú vysoko korelujúce s cenou Bitcoinu.

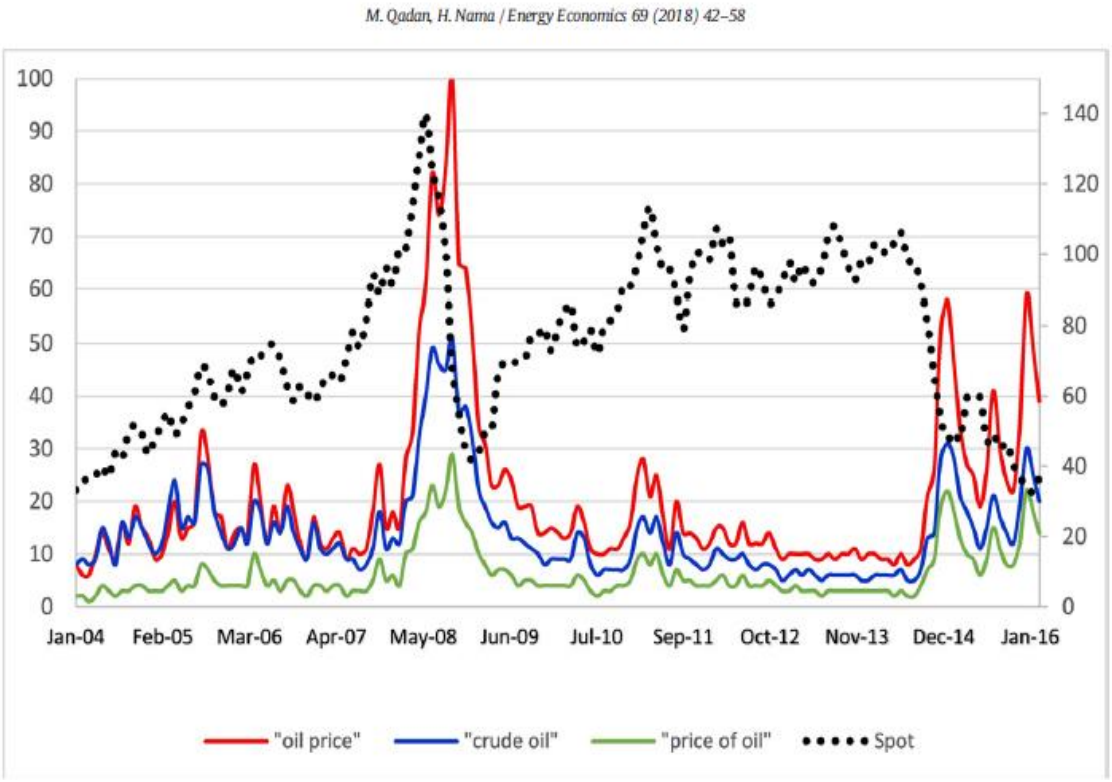
Obrázok 2. Graf vývoja cien Bitcoinu za obdobie sledované autormi [8]

2.2.2. Trendy vo vyhľadávaní

Jedným z atribútov využívajúcich vplyv sentimentu, boli použité i dáta z Google Trends, ktoré vyjadrujú mieru záujmu verejnosti o zvolenú problematiku. V praxi si môžeme pod týmto termínom predstaviť frekvenciu, prípadne množstvo vyhľadávaní termínu blockchain a Bitcoin na Google vyhľadávači. Abraham, Jethin et al. sa v práci [9] zamerali na predikciu ceny Bitcoinu na základe dát z Google Trends. Samotné dáta vyhodnotili ako vysoko korelujúce s vývojom cien. Trendy vo vyhľadávaní autor kombinoval s množstvom zverejnených správ cez sociálnu sieť Twitter, čo hodnotíme ako nie úplne najšťastnejšiu možnosť, nakoľko informácia o miere záujmu verejnosti, obsahujú práve dáta z Google Trends a objem správ na sociálnej sieti Twitter tvorí len menšiu podmnožinu celkového záujmu verejnosti.

Celkovo však autor otvára možnosti širšieho zahrnutia názoru verejnosti na danú doménu. Tieto pozorovania môžu výrazne prispieť k lepšiemu vyprofilovaniu verejnej mienky o stave daného trhu (trhového sentimentu). Zároveň však nemôžme opomenúť vhodnosť využitia sociálnych sietí ako zdroja vhodných informácií. Namiesto počtu odoslaných správ v oblasti danej domény by sme zvolili prístup analýzy obsahu daných správ, nakoľko čistý objem nemusí priamo určovať kladný alebo záporný vývoj trhu.

Okrem kryptomien existujú i iné trhové segmenty, ktoré môžu výraznou mierou čerpať z verejného sentimentu. Autor Mahmoud Qadan sa vo svojej práci [10] pokúsil vysvetliť úzke prepojenie medzi cenou ropy a mierou vyhľadávania a sentimentu z Google Trends. Počas celej práce autor konštatuje vysoké korelácie medzi sledovanými atribútmi a finálnym atribútom - cenou. Využíva k tomu 9 atribútov zahŕňajúc práve Google Trends, týždenné a mesačné publikácie (už od roku 1973) a rôzne stresové indexy. Výsledkom jeho práce je graf znázornený nižšie (obrázok 3) a prezentuje vysokú podobnosť medzi mierou sentimentu týchto publikácii a ceny ropy.



Obrázok 3. Graf vývoja jednotlivých vyhľadávaných termínov a cenou ropy. Obrázok prevzatý z práce [10]

Výsledky tohto autora nám môžu pomôcť pri výbere jednotlivých zdrojov pre extrakciu sentimentu. Odlišnosti sú však markantné, nakoľko produkt ako ropa má dlhšiu históriu a spôsoby šírenia informácií bol výrazne obmedzenejší a pomalší. Kryptomeny majú však základ v technickom zázemí. Ide o digitálny prostriedok, ktorý zaznamenal extrémnu vlnu mediálneho záujmu a široký záujem verejnosti. Preto by nemalo byť ťažké získať dostatok zdrojov pre extrakciu sentimentu. V tomto momente však narážame na rozdiel oproti práci [10], nakoľko jeho historické zdroje, pochádzajúce z dávnejšej histórie, tvorili žurnály a články, ktoré predstavujú autoritu v danom odvetví. Nakoľko história kryptomien je relatívne krátka, problém môže spočívať pri výbere relevantných a názorotvorných

publikácií. Našou snahou bude vyhnúť sa publikáciám zo stránok so slabým verejným zázemím, prípade neoznačených blogov. Uprednostňované budú zdroje, ktoré nesú informáciu o miere prijatia danej správy verejnosťou, napríklad páčiky (z angl. likes) a zdieľania (z angl. shares), čím by sme mali eliminovať zašumenie vstupných dát autormi prezentujúcimi vlastné, ničím nepodložené názory a teórie.

2.2.3. Sentiment

Okrem historických dát sa viaceré štúdie, ako sme už naznačili vyššie, venujú schopnosti predikcie vývoja cien zo správania sa spoločnosti na internete [4], [9]–[11]. Celkovo by sme si pod pojmom sentiment mohli predstaviť náladu spoločnosti a jej zmýšľanie. Tento fakt môže výrazne pomôcť pri určovaní smeru vývoja danej komodity. V závislosti od kladného či záporného ladenia správ, dokážeme odhadnúť nárast, či pokles vývoja cien. Tiež musíme brať ohľad na fakt, že väčšina správ vzniká ako reakcia na stav trhu. Tieto správy majú silný predpoklad odchytiť prvopočiatok nárastu alebo pádu cien. Hovoríme o davovom efekte na sociálnych sieťach, kedy výrazne negatívne alebo pozitívne ladené správy môžu mať efekt na väčšiu skupinu obchodníkov, čím prehĺbia efekt, ktorý sa na trhu vyskytol. Taktiež môžeme odchytiť správy autorít v danom odvetví, nakoľko viacero účastníkov daného trhu sú silne fixovaní na vyjadrenia odborníkov a reagujú na akékoľvek správy impulzívne.

Pri procese extrakcie sentimentu silne závisí od kontextu jednotlivých správ. Nie je vôbec ojedinelé, kedy o miere sentimentu jednotlivých slov rozhoduje až samotný kontext. V tomto prípade sa autori opierajú o extrakciu sentimentu na úrovni slov a n-tíc slov takzvaných n-gramov. Extrakcia n-gramov zo správ môže výrazne zvýšiť presnosť určenia správnej miery sentimentu. Taktiež je veľmi dôležité poznať prostredie, v ktorom dané správy vznikajú. Rôzne prístupy musia byť aplikované v prípade ak ide o blogy, kde autori prezentujú svoj názor štruktúrovanou formou, či správ uverejňovaných na fórach alebo prostredníctvom sociálnych sietí. Sociálne siete v tomto prípade tvoria špeciálnu sekciu, nakoľko používatelia využívajú pri komunikácií zvyčajne slang, skratky, či vyjadrujú svoje pocity a emócie emotikonmi. Obzvlášť pri sociálnej sieti Twitter je dôležité spomenúť, že tieto správy sú obmedzené dĺžkou na 280 znakov.

Connor Lamon sa vo svojej práci [4] zameriava na vplyv sentimentu z webových článkov a sociálnych sietí na cenu kryptomien. Zameriava sa celkovo na tri kryptomeny a to konkrétne na Bitcoin, Litecoin a Ethereum. Jeho historické dáta pochádzajú z časového spektra 67 dní. Taktiež na začiatku svojej práce zadefinoval dve základné otázky:

- môže analýza sentimentu webových článkov prípadne postov zo sociálnych sietí poskytnúť dostatočnú výpovednú hodnotu pre úspešnú predikciu ceny kryptomien?
- V prípade pozitívneho zistenia, ktorý zo zdrojov sa zdá byť ako lepší indikátor budúcej ceny? [4]

Cieľom autorov je implementácia zozbieraných poznatkov do širšieho inteligentného systému, ktorý bude spravovať ich portfólio v oblasti kryptomien. Ďalej uvádzajú, ako výhodu svojho riešenia spojenie dvoch nezávislých zdrojov dát pre analýzu sentimentu a taktiež zameranie sa rovno na tri kryptomeny čím diverzifikuje možné riziko. Zároveň uvádzajú, že ich model bude vyhodnocovaný na základe percentuálnej zmeny, v čom vidia nedostatok predchádzajúcich štúdii dosahujúcich 80 a viac percentné úspešnosti modelov. Autori celkovo pracujú nad dátami z obdobia nasledujúcich objemov:

- 3600 článkov o kryptomenách,
- 10 000 tweetov ohľadom Bitcoinu,
- 10 000 tweetov ohľadom Litecoinu,
- 10 000 tweetov ohľadom Etherea.

Jednou zo slabých stránok prác [4], [12] je, že historické dáta (cenu) využívajú iba na označkovanie svojich dát, inak s nimi pri trénovaní svojho modelu nepočítajú. Je veľmi ťažké validovať kvalitu vstupných dát, nakoľko autori neuvádzajú časové spektrum, nad ktorým bola ich práca vypracovaná. Pri značkovaní pritom využívajú princíp, kedy každému článku prípadne tweetu doplnia cieľové atribúty (2 pre každú menu), kde prvý predstavuje cenu v dni $t + 1$ a druhý predstavuje cenu v dni $t + 2$. Samotná značka má binárnu hodnotu 0/1 pre pokles prípadne nárast danej kryptomeny. Predspracovanie textu zahŕňa odstránenie bielych znakov, diakritiky, stop znakov a konverziu celého textu na malé písmená. Následne sú skonštruované 1 a 2-gramy pre lepšie výsledky analýzu textu.

Hoci autori riešia binárny klasifikačný problém, metriku vyhodnocovania používajú s ohľadom na zmenu ceny danej kryptomeny. Výsledná metrika je porovnanie priemernej ceny pri správnych predikciách s priemernou sumou nesprávnych predikcií. Autori taktiež využívajú viacero modelov (tabuľka 2) pri predikcií a vo výsledku uvádzajú ku každej mene najlepší model a zvolený prístup. Celkovo však hodnotia najlepšie model lineárnej regresie, pri ktorej dosahujú najlepšie výsledky.

2.3. Prehľad výsledkov príbuzných prác

V tabuľke 2 môžeme vidieť výsledky jednotlivých autorov zaoberajúcich sa obdobnou tematikou. Ich výsledkami sú hodnoty rôznych vyhodnocovacích metrík, ktoré vďaka svojim rozličným prístupom nie je možné navzájom efektívne porovnať.

Tabuľka 2. Zobrazenie prehľadu dosiahnutých výsledkov s použitými metódami

Autor	Dáta	Prístup	Cieľ	Výsledok
[6]	Historické dáta Fundamentálne dáta	RNN, LSTM, ARIMA	Nárast a pokles budúcej ceny	Úspešnosť 52%
[11]	Twitter dáta	LR, SVR, Naivný Bayes	Nárast a pokles v časovom okne (hodina a deň)	Úspešnosť 73.5 % a 91.4%
[4]	Twitter dáta	Logistická regresia, Naivný Bayes, SVM	Nárast a pokles budúcej ceny	Úspešnosť 52-60%
[9]	Twitter dáta, Google Trends	F skóre P- hodnota	Miera korelácie	Korelácia 0.817 0.841
[13]	DJIA Google Trens Twitter dáta,	Štatistické modely	Cena budúcich období	MAPE 0.253-4.148
[14]	Twitter dáta	Analýza korelácií	Miera korelácie	Jednotlivé atribúty 0.212-0.896
[15]	Komentáre a Zdieľania	Vlastný štatistický model	Nárast a pokles v daný deň	Presnosť 74%
[12]	Historické dáta	Vlastný model	Cena v daný deň	RSME 0.075
[7]	Historické dáta	RNN, LSTM	Nárast a pokles budúcej ceny	Úspešnosť 52%
[8]	Historické dáta	Lineárna regresia Náhodný les Gradient Descent	Nárast a pokles v daný deň	Úspešnosť 99.4%
[16]	Online články	ARIMA	Nárast a pokles v daný deň	94.3%
[17]	Historické dáta Blockchain dáta	Binomický GLM Náhodný les SVM	Nárast a pokles v daný deň	Úspešnosť 98.79%

2.4. Použité metódy

Autori vo svojich prácach pristupovali k danému problému rôznymi spôsobmi. Okrem rôznych vyhodnocovacích metrík, autori svoje výsledky dosiahli rôznymi metódami. Analýzou ich riešení sme získali prehľad o viacerých prístupoch, kde možno všetkých autorov podľa prístup zatriediť do troch nasledujúcich skupín:

- strojové učenie,
- neurónové siete,
- štatistické modely.

2.4.1. Strojové učenie

Predstavuje podskupinu umelej inteligencie, ktorá je zameraná na aplikáciu metód a algoritmov umožňujúcimi danému modelu učiť sa. Tento model vie následne adekvátne reagovať na rôzne vstupné hodnoty, bez predošlej dodatočnej informácie alebo explicitnému programovému zásahu. Algoritmy strojového učenia čerpajú poznatky zo štatistickej analýzy, matematickej štatistiky a hĺbkovej analýzy dát. Výsledkom daného algoritmu je model, vykazujúci určitú vedomosť, pri vstupných parametroch bez nutnosti ich predošlej znalosti. Celkovo môžeme metódy strojového učenia rozdeliť do troch skupín:

- učenie s učiteľom (z angl. Supervised learning),
- učenie bez učiteľa (z angl. Unsupervised learning),
- učenie s posilňovaním (angl. Reinforcement learning) [20].

Tieto prístupy riešia viaceré typy úloh, ktoré sa môžu vyskytnúť. Napríklad pre klasifikačné úlohy so známymi cieľovými atribútmi je najbežnejším prístupom učenie s učiteľom, nakoľko sa snažíme klasifikovať do vopred známych tried. V prípade ak cieľové atribúty nepoznáme, môžeme zvoliť prístup učenia bez učiteľa, kde model bude rozhodovať na základe podobností a objavuje takzvané zhluky dát, ktoré sú si niečím podobné. Ďalším prípadom môže byť regresná úloha, ktorá bude aj našim predmetom skúmania. Jej cieľom je odhadnúť presný budúci vývoj nejakej situácie, napríklad počet zdieľaní nového videoklipu, prípadne veľkosť zrážok v jednotlivých dňoch. Vyššie spomenutí autori zaoberajúci sa problémom predikcie cien Bitcoinu, využívali vo svojich prácach dva najzákladnejšie prístupy. Prvým bola binárna klasifikácia, kde predikovali budúci nárast alebo pokles ceny a druhým bola regresia, kde sa snažili presne vyjadriť mieru nárastu alebo poklesu ceny. Z hľadiska využiteľnosti je logickejšou voľbou regresný prístup, nakoľko pri binárnej klasifikácií pri nezohľadnení miery nárastu alebo poklesu riskujeme zlú, prípadne

zavádzajúcu interpretáciu výsledkov. Nižšie spomenieme niektoré modely, ktoré boli vybrané jednotlivými autormi a sú spomenuté v tabuľke 2.

Lineárna regresia je najjednoduchší algoritmus regresnej analýzy využitý v prácach autorov [8], [11]. Na rozdiel od ostatných algoritmov, ktoré patria ku klasifikačnému problému a vytvárajú nespojitý výstup, lineárna regresia patrí k regresnému problému a jej výsledkom je spojitý výstup. Regresia vie určiť závislosť medzi vstupom a výstupom z tréningových dát a následne vie vypočítať závislú premennú, teda výstup. Jej výpočet je realizovaný pomocou nasledujúcej rovnice:

$$y = ax + b$$

Rovnica 1. Vzťah pre výpočet závislej premennej v lineárnej regresií

Pre Rovnica 1. Vzťah pre výpočet závislej premennej v lineárnej regresií platí, že:

- **y** je výstup lineárne regresie, teda závislá premenná,
- **x** je vstup, teda nezávislá premenná, od ktorej závisí výstup,
- **a** je prvok zvaný slope a vyjadruje sklon regresnej priamky na bodovom diagrame. Slope môžeme eventuálne nazvať aj ako rozostup medzi dátovými bodmi. Tento prvok zostáva pri lineárnej regresii rovnaký.
- **b** je prvok zvaný intercept. Vyjadruje odskok začiatku regresnej priamky od nulového bodu. Ak je intercept rovný 0, tak sa regresná priamka bude začínať v nulovom bode. Ak je intercept napríklad 1, regresná priamka bude začínať v bode so súradnicami [1;1] [21].

Naivný Bayes je algoritmus využitý v práci [4], [11], ktorý patrí medzi takzvané pravdepodobnostné klasifikátory a patrí medzi tie najjednoduchšie. Najčastejšie sa využíva na prácu s textom, ako napríklad identifikovanie spamu alebo zaradenie článku podľa jeho obsahu. Pri algoritme Naive Bayes sa pravdepodobnosť počíta cez Bayesovu vetu. Táto veta je matematicky vyjadrená rovnicou:

$$\mathrm{P}(\mathrm{A}|B) = \frac{P(B|A)P(A)}{P(B)}$$

Rovnica 2. Vzťah pre Bayesovu vetu

kde:

- P(A|B) je výsledná pravdepodobnosť. Tú sa model snaží vypočítať a podľa toho, pri ktorom objekte je vyššia, určuje ku ktorému označeniu patria dáta.

- A je hypotéza, ktorá je daná javom ,
- B, ktorý slúži ako dôkaz, ktorý potvrdzuje hypotézu A.
- P(B|A) je pravdepodobnosť dôkazu B, za predpokladu, že hypotéza A je pravdivá.
- P(A) je pravdepodobnosť, že hypotéza A je pravdivá, bez ohľadu na dôkaz B.
- P(B) je pravdepodobnosť, že dôkaz B je pravdivý, bez ohľadu na hypotézu A [22].

Rozhodovacie stromy sú zložené z uzlov, pričom základný uzol sa nazýva koreňový uzol. Využívané boli autormi [8], [11], [17]. Uzly sa ďalej rozvetvujú a vytvárajú tak štruktúru stromu. Každý uzol predstavuje rozhodovanie podľa jednej vybranej vlastnosti klasifikovaného objektu. Vyberané vlastnosti musia objekty od seba čo najviac odlišovať, aby boli na konci stromu čo najpresnejšie klasifikované. Pre správny výber atribútu, ktorý musí čo najlepšie rozdeliť dáta do dvoch čo najodlišnejších vetiev, sa využíva informačná entropia a tzv. informačný zisk. Vzorec entropie je nasledovný:

$$h = -\sum_{i}^{n} (Pi) log2(Pi)$$

Rovnica 3. Vzťah pre výpočet entropie

kde:

Pi je zlomok príkladov v triede i v trénovanej podmnožine dátovej množiny.

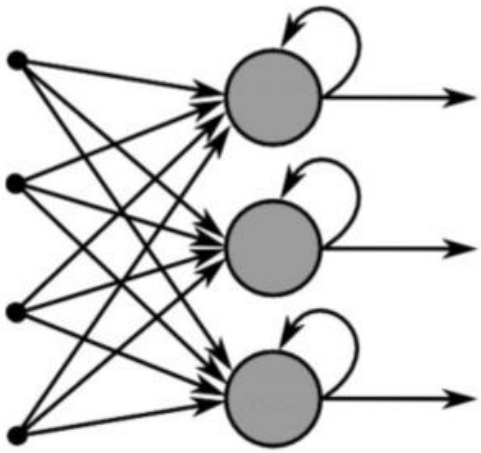
Ak je entropia pri použití log2 rovná nule, potom všetky príklady sú tej istej triedy. Ak je entropia pri použití rovnakého logaritmu rovná jednej, čo je jej maximálna hodnota pri tomto logaritme, príklady sú rovnomerne rozdeliteľné [23].

2.4.2. Neurónové siete

Boli použité v prácach autorov [6], [7], [16], [17] a sú podľa [25] definované ako masívny paralelný výpočtový systém, ktorý má schopnosť uchovávania informácii a umožňuje ich ďalšie spracovanie, pričom inšpiráciu čerpá z funkcií ľudského mozgu. Základným stavebným prvkom neurónovej siete perceptrón. Samotný neurón však vykonáva operáciu, ktorá transformuje vstup na výstup vo veľmi jednoduchej podobe. K docieleniu komplexnejších riešení sa využíva princíp spájania týchto neurónov do takzvaných vrstiev. Tieto vrstvy potom v celku predstavujú výslednú sieť neurónov [25].

Neurónové siete podľa svojej stavby môžu plniť špecifické úlohy. Ako je aj uvedené v tabuľke 2 viacerí autori pracujú s konceptom neurónových sietí. Výrazne vystupuje iba jeden typ neurónových sietí a to konkrétne rekurentné neurónové siete (RNN).

RNN predstavuje typ neurónovej siete, ktorá rieši problém zlyhávania bežných nelineárnych viacvrstvových sietí v dôsledku sledovania časovej závislosti. Celkovo tento typ siete rieši problém, keď časový rad tvorí vstup pre danú sieť, ktorá musí reagovať na atribút tohto typu. Riešením je vytvorenie vstupu, ktorý je generovaný niektorými neurónmi a následne je spracovávaný v sieti opakovane. Tým dochádza k zohľadňovaniu predchádzajúcej informácie, ktorá cez sieť prešla [25]. Schému RNN môžeme vidieť na obrázku 4.



Obrázok 4. Schéma rekurentnej neurónovej siete

Hlavnou nevýhodou neurónových sietí je problém pri ich interpretácií. V prípade, ak by sme chceli vyhodnotiť proces správania sa modelu a založiť na ich základe ďalšiu etapu výskumu, napríklad optimalizáciu obchodných pravidiel, v prípade neurónových sietí narážame problém interpretovateľnosti. Naopak pri dosahovaní finálnych údajov neurónové siete s dostatkom vstupných dát a dostatočným počtom záznamov dokážu dosahovať lepšie výsledky.

2.4.3. Štatistické modely

Niektorí autori [13], [15], [12] sa pokúšali vytvoriť vlastné modely založené na štatistických metódach. Ich hlavnou motiváciou bola úzka koncentrácia sa na vybrané časti sledovaných čŕt, pri ktorých sú už známe postupy efektívneho predikovania. Avšak ich úzka špecializácia na vybrané časti celkového problému tvorí ťažko overiteľný a reprodukovateľný prístup. Taktiež považujeme za nevhodné explicitné uvedenie vzťahov výpočtu nakoľko týmto prístupom sa z modelu stáva model neschopný reagovať na nové skutočnosti (nové atribúty, rozloženia atribútov,..). Z týchto dôvodov túto možnosť ani nebudeme ďalej uvažovať.

2.5. Zber dát

Nakoľko už v priebehu tejto analýzy máme hotový prototyp zberu dát zo sociálnej siete Twitter, môžeme porovnať objemy autorov [4], ktoré pre svoju predikciu používajú. Vyššie bolo spomenuté, že autori pracujú v časovom spektre 67 dní. Pri obnose 10 000 tweetov autori majú k dispozícií priemerne 150 tweetov za deň, čo považujeme za extrémne nízku hodnotu. Dôvod bude asi v spôsobe zberu dát, kde sa zamerali na stránky poskytujúce informácie v oblasti kryptomien, čím obmedzili vplyv širšej verejnosti.

Naše riešenie zberu dát je v prevádzke od 26.03.2018 a k dňu 15.05.2018 sme celkovo zozbierali obnos tweetov obsahujúce slová „Bitcoin“ a „Btc“ o veľkosti takmer 4GB. Štatistiku získaných údajov je možné vidieť v tabuľke 3. Ako je možné vidieť reálny objem správ o Bitcoine je výrazne väčší. Taktiež treba podotknúť, že v našej dátovej množine sú dáta surové a nespracované. To znamená, že obsahujú výrazné množstvo duplicít, ktoré vznikli zdieľaním týchto správ. Autori [11], [13], [14] vo svojich prácach tieto duplicity odstraňujú. Tento krok však považujeme za rozporuplný, pretože autori prichádzajú o informáciu obľúbenosti resp. stotožňovania sa s danou informáciou. Užívateľ, ktorý preposiela rovnakú správu ako niekto predtým, vyjadruje určitú sympatiu s týmto názorom a prístupom odstránenia duplicít sa táto informácia vytráca. Navrhujeme prístup, kedy sa duplicity odstránia, ale pred samotným odstránením bude zistený počet identických správ, čím vytvoríme váhu danej správy.

Tabuľka 3. Sumárny prehľad získavanej dátovej množiny

Názov hodnoty	Hodnota
Maximálny počet tweetov za deň	152442
Minimálny počet tweetov za deň	72731
Priemerý počet tweetov za deň	85456
Stredná hodnota počtu tweetov za deň	81525

3. Zhrnutie analýzy

Analýzou obdobných prác autorov sa nám vyprofilovali hlavné smery, ktoré môžu byť v našej práci ťažiskové. Dôležitým výsledkom hĺbkovej analýzy je fakt, ktorý potvrdzuje úzku previazanosť verejnej mienky ale aj technických atribútov. Ako najdôležitejšie faktory ovplyvňujúce vývoj cien kryptomien sa javia historické dáta, či už o vývoji ceny alebo stavu blockchainu, sentimentu a v miernom prípade i vývoji ostatných kryptomien.

Sentiment dokáže byť extrahovaný z viacerých portfólií, nakoľko kryptomeny ako komodita sú extrémne mladá súčasť trhovej ekonomiky, no napriek tomu vzbudila nemalý záujem verejnosti. Náladu verejnosti preto nebude tak náročné vyprofilovať ako v prípade iných druhov komodít, nakoľko téma kryptomien je obľúbená a počas svojho rozkvetu vznikli mnohé fóra zaoberajúce sa výhradne touto tematikou. Taktiež existujú blogy, ktoré poskytujú štruktúrované články na túto tému a dokážu byť strojovo spracované. Dobrým poznatkom je aj fakt, že autori sa zameriavajú aj na globálnu mienku a nekladú dôraz len na autority v danom odvetví. Tento fakt môže byť zavádzajúci a môže vyžadovať viacero experimentov, kde by bolo vhodné preukázať, či verejnosť ako taká a jej nálada dokáže profilovať cenu kryptomien alebo sú to skôr články v odborných publikáciách, či názory a blogy autorít v danom odvetví. Pre verejný názor a ladenie spoločnosti využijeme pravdepodobne sociálnu sieť Twitter, či Google Trends a v prípade názorov autorít to budú publikované články na rôznych fórach.

V rámci extrakcie sentimentu musia byť použité taktiež odlišné techniky. Pri dolovaní sentimentu zo štruktúrovaných textov nám pomôžu zavedené slovníky a nástroje. V prípade tweetov zo sociálnej siete budeme musieť použiť sofistikovanejšie nástroje, nakoľko správy bývajú častokrát neštruktúrované, prípadne využívajú aj iné ako slovné vyjadrenia emócií napríklad emotikony, čo môže byť výrazným uľahčením pri extrakcií sentimentu z krátkych neštruktúrovaných príspevkov.

Našou snahou bude obsiahnuť čo najväčšie spektrum spoločnosti, nakoľko nám ide o odchytenie jednak globálnej verejnej mienky, čo môžu byť napríklad tweety prípadne dáta z Google Trends ale aj publikácie rôznych autorít v danej doméne, či online verzie tlačených publikácií.

Nemenej dôležitými atribútmi budú historické dáta. Našim cieľovým atribútom bude trhová cena Bitcoinu. Samotná blockchain technológia nám poskytuje podrobné záznamy o niektorých veľmi úzko súvisiacich atribútoch. Tieto dáta nebude ťažké získať, nakoľko vo väčšine prípadov ide o verejne dostupné informácie.

Jednotliví autori dosahujú široké spektrum úspešností, čo má za následok hlavne rôzne vyhodnocovacie metriky zvolených autorov a definíciu cieľového atribútu. Najlepšie výsledky dosahujú autori s binárnymi klasifikáciami, ktoré však neodzrkadľujú mieru a závažnosť (výnosnosť prípadne stratu) danej predikcie. Tento fakt by spôsobil mnohým autorom problémy, napríklad pri použití agenta využívajúci ich model ako nástroj na správu finančného portfólia v praxi. Hlavným dôvodom je nevhodné využívanie metriky úspešnosť, ktorá je vo finančnej sfére nevhodná, nakoľko pre potenciálneho investora má väčší negatívny vplyv správa chybného nákupu ako správa chybného predaja. Z tohto dôvodu bude binárna klasifikácia prevedená na úrovni dodatočného získania informácií o smere pohybu Bitcoinu.

Ďalším dôležitým aspektom je sledované obdobie daného trhu. Autori dosahujúci vysokú úroveň úspešnosti uskutočňovali svoj výskum na výrazne malej vzorke dát. Ďalším dôvodom vysokej úspešnosti môže byť vplyv nevhodne zvoleného cieľového atribútu. Zaujímavejším prístupom môže byť väčšie časové obdobie, ktoré zaznamenáva rovnomerné pohyby danej meny (nie čisto stúpajúcej alebo klesajúcej periódy). To je ďalší fakt, ktorí mnohí autori svojich prác neuviedli explicitne a ich výsledky môžu byť touto skutočnosťou výrazne ovplyvnené. Ide o princíp kedy sledované obdobie obsahuje takmer výlučne dáta stúpajúceho (klesajúceho) charakteru. V tom prípade modely len prehlasujú výsledok na základe početnejšej triedy, čím strácajú svoju predikčnú schopnosť.

Z toho dôvodu sme sa rozhodli zozbierať údaje z väčšieho časového spektra rádovo v rámci mesiacov. Tento prístup má väčší predpoklad, že zachytíme oba smery vývoja danej kryptomeny. Ako sme už vyššie spomenuli mechanizmus na zber dát zo sociálnej siete Twitter je nasadený na externom serveri a zbiera dáta od 26.03.2019. V prípade, ak nenastanú žiadne komplikácie, by sme mali pri vyhodnocovaní disponovať dátami z takmer celoročného obdobia, čo bude časové obdobie, ktoré žiaden z vyššie spomenutých autorov neskúmal.

4. Opis riešenia

Proces riešenia práce prebiehal vo viacerých fázach. Nakoľko sme v predchádzajúcich častiach definovali okruhy dát, na základe ktorých sa autori pokúšali predikovať vývoj cenovej hladiny Bitcoinu, našou prvou úlohou bolo zozbierať dostatočné množstvo z čo najširšieho okruhu. Hlavnou motiváciou je čo najlepšie pokrytie faktorov, ktoré môžu ovplyvňovať vývoj výslednej ceny. Ako sme poznamenali v časti analýzy, častým problémom bol práve nedostatok dát, prípadne z nášho pohľadu veľmi krátke sledované obdobie. Tento problém sme sa snažili eliminovať na začiatku našej analýzy kedy bol spustený skript zabezpečujúci zbieranie dát zo sociálnej siete Twitter. Práve tieto dáta sa stávali úzkym hrdlom viacerých prác, nakoľko Twitter neposkytuje historické dáta verejne ale poskytuje živú API službu pomocou ktorej sa dajú tieto dáta zbierať. Z toho dôvodu všetky naše dáta budú sledované od 26.03.2019 kedy bol skript nasadený na server a boli sme schopní tieto údaje zbierať. Finálnym cieľom je však vybudovať dátovú množinu, ktorá bude obsahovať široké spektrum čŕt, ktoré v teoretickej rovine môžu pomôcť pri predikcií. Celkovo sme sa zamerali na niekoľko skupín, ktoré spĺňajú tento predpoklad. Tieto skupiny sú:

- **Twitter dáta** – všetky správy (tweety) odoslané na sociálnej sieti Twitter, ktoré obsahujú #bitcoin, #BTC,
- **články a fóra** – publikované články na portáloch zameriavajúcich sa na trh kryptomien alebo Bitcoin špeciálne,
- **dáta Bitcoinu** – historické údaje o Bitcoine s údajmi blockchainu,
- **trh akcií** – historické údaje o akciách spoločností v IT sektoroch a tzv. pilierové komodity (ropa, zlato,..),
- **trh kryptomien** – historické dáta viacerých známejších kryptomien (Ethereum, Litecoin, Monoro,..).

V ďalších kapitolách sa zameriame na proces získavania, štruktúru a hĺbkovú analýzu jednotlivých podskupín s dôrazom na možnú reprodukovateľnosť našich výsledkov. V hĺbkovej analýze sa pozrieme na rozdelenie dát a proces ich úpravy do podoby vyhovujúcej pre budúcu predikciu.

4.1. Časové rady a nimi definované problémy

Časový rad tvoria hodnoty, ktoré sú sledované v časovom slede a ich hodnota je od časového aspektu závislá. Čas však nie je jediná nezávislá premenná, od ktorej vývoj daného časového radu závisí [26]. V našej práci sa budeme stretávať primárne s diskrétnymi časovými radmi,

pre ktoré platí konštantná časová zložka medzi jednotlivými pozorovaniami (minúta, hodina, deň,..). Vzhľadom na vyššie spomenuté aspekty časových radov je zrejmé, že proces analýzy nebude rovnaký ako pri analýze nezávislých pozorovaní. S časovými radmi sú definované problémy, ktoré v základe od situácie a zvoleného postupu je nutné riešiť. Nižšie vyberieme najčastejšie sa vyskytujúce problémy.

4.1.1. Problém multikolinearity

Problém multikolinearity je častým javom pri riešení regresných problémov. V našom prípade tento problém výrazne vystupuje do popredia, nakoľko multikolinearita je definovaná podľa [27] ako závislosť jednotlivých vstupných premenných navzájom, čím sa porušuje predpoklad nezávislých vstupných premenných. To pri sledovaní cien jednotlivých ekonomík môže byť problém, nakoľko jednotlivé skryté korelácie môžu byť ťažko odhaliteľné, či dokonca náhodné. Miera závažnosti tohto problému je striktne závislá od konceptu v ktorom vystupuje. Jej detekcia je však v niektorých prípadoch netriviálny problém. V tabuľke 4. uvádzame jednoduchý príklad multikolinearity.

Tabuľka 4. Príklad multikolinearity dát

Počet rokov praxe	Vek kandidáta	Úroveň seniority (0-100)
2	28	30
8	45	47
15	50	70
22	47	78

Ako je zrejmé, problém v tomto príklade sú atribúty počet rokov praxe s vekom kandidáta. V praxi to znamená, že je náročné zmeniť hodnotu atribútu bez vplyvu na ostatné. V tomto prípade, ak sledujeme rovnakého záujemcu v čase, je zrejmé, že ak počet rokov praxe stúpa, automaticky stúpa aj vek daného kandidáta. Postup detekcie multikolinearity spočíva v zistení korelácií jednotlivých dvojíc a hĺbková analýza závislých dvojíc. Proces riešenia zvyčajne spočíva vo vybraní signifikantnejšej črty, čo by v našom prípade predstavoval atribút počet rokov praxe.

4.1.2. Problém autokorelácie

Problém autokorelácie je sledovateľný pri časových výraznejšie ako pri bežných - nezávislých pozorovaniach. Hlavným problémom je práve vývoj jednotlivých atribútov v čase za predpokladu, že minulé pozorovanie, prípadne pozorovania a jeho hodnoty sú závislé a minulá a hodnota v čase t je výrazne podobná hodnote v čase $t-1$. V závislosti od posunu času t o k časových jednotiek môžeme hovoriť o autokorelácií k-teho stupňa

[27]. Miera autokorelácie sa dá merať Durbin – Watsnovým testom pre mieru autokorelácie. Vzťah je definovaný ako:

$$\mathrm{d} = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{t} e_t^2}$$

Rovnica 4. Vzťah pre výpočet Durbin Watson testu

kde d nadobúda hodnoty z intervalu $< 0,4 >$ pričom interval $< 0,2)$ predstavuje mieru pozitívnej korelácie $(2,4 >$ predstavuje mieru negatívnej korelácie, pričom hodnoty v blízkom okolí 2 sú považované za nekorelované v zmysle autokorelácie. Postupov odstránenia efektu autokorelácie je niekoľko. Všeobecne najzaužívanejší je však proces diferenciácie kedy je z daného časového radu extrahovaná zmena v jednotlivých dňoch. Všeobecne môžeme diferenciu d definovať ako :

$$d = X_t - X_{t-1}$$

Rovnica 5. Vzťah pre výpočet diferencie hodnoty X

Kde d je diferencia danej hodnoty a X_t je sledovaný atribút v čase t. Aplikáciou vyššie uvedeného vzťahu získame $n - 1$ atribútov, ktoré vyjadrujú mieru zmeny hodnoty oproti minulému pozorovaniu s absenciou vplyvu „rovnakej“ zložky. Týmto prístupom však strácame informáciu o pôvodnej hodnote daného atribútu, preto sa často využíva aj vzťah:

$$md = X_t - priemer(X)$$

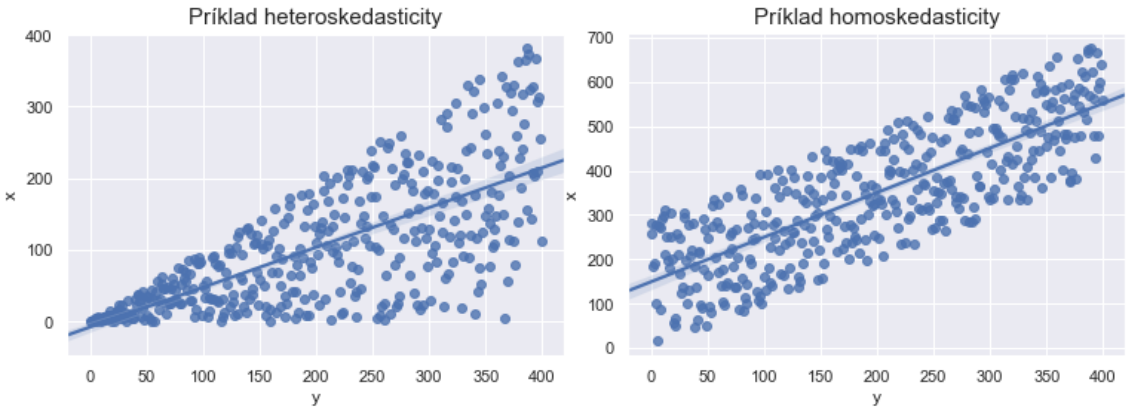
Rovnica 6. Variácia výpočtu diferencie

Tento vzťah zmierňuje vplyv autokorelácie spolu s jednoduchým procesom spätného výpočtu danej hodnoty nakoľko priemer(X) je vzhľadom na statickú množinu pozorovaní konštantný.

4.1.3. Problém heteroskedasticity

Tento druh problému síce nesúvisí priamo s časovými radmi, no je častokrát pozorovaný vo všetkých typoch regresných problémov. Heteroskedasticita predstavuje mieru nekonštantnosti rozptylu [28]. Na obrázku 5 môžeme vidieť heteroskedastický a homoskedastický model rozloženia dát, pričom práve homoskedastické rozloženie je žiadúce. Proces detekcie je prevádzaný zvyčajne prvotným vykreslením sledovanej hodnoty, ktorá je následne preverená White testom. Tento test meria konštantnosť variácie a jej zachovanie na celom rozsahu daného atribútu. Transformácia heteroskedasticitského

rozloženia na homoskedasticitské spočíva vo väčšine prípadov v aplikácií logaritmickej funkcie, ktorá zjemní vplyv výkyvov vzdialenejších hodnôt.



Obrázok 5. Príklad heteroskadistického a homoskadistického rozloženia dát

4.2. Dolovanie a analýza historických dát Bitcoinu

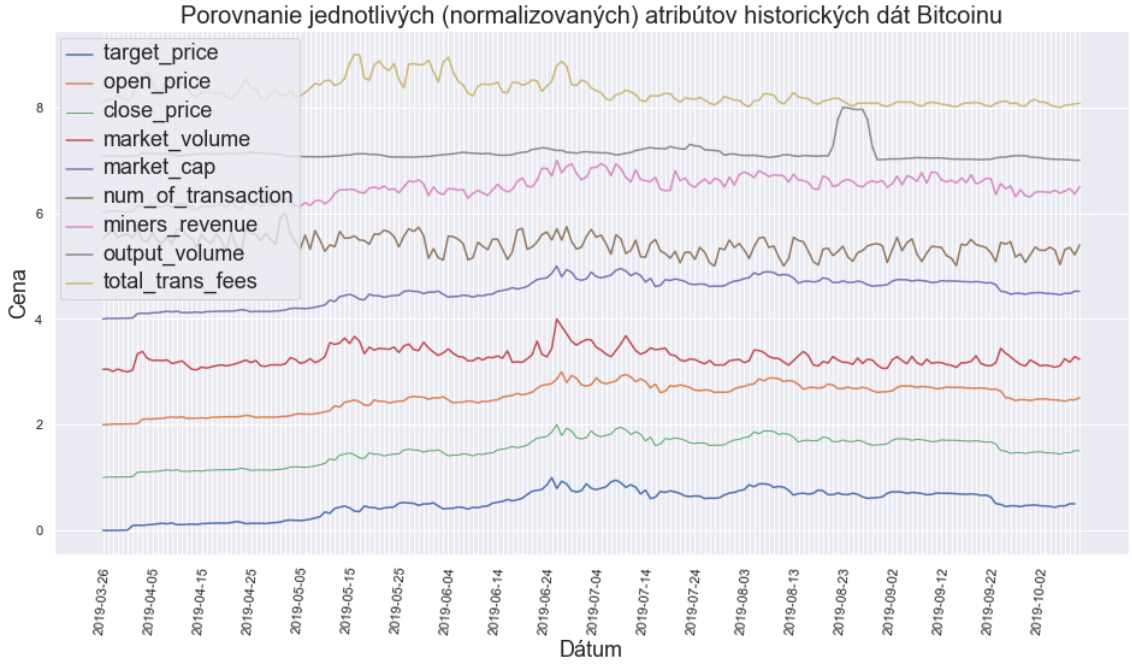
Pre pochopenie problematiky a položenie základu našich dát sme si vybrali historické dáta Bitcoinu. Dôvod je hlavne ten, že práve táto skupina obsahuje náš budúci cieľový atribút – cenu. Tak ako v mnohých prípadoch, tak aj Bitcoin obsahuje historickú stopu, ktorú zanecháva pri svojom pôsobení. Nejedná sa však len o črty samotnej kryptomeny ale aj technológie na základe ktorej vznikol – blockchain. Touto skupinou dát pokryjeme historické hodnoty jednotlivých čŕt priamo súvisiacich s danou kryptomenou. Jednotlivé dáta boli získavané z online zdrojov [29]. Nakoľko tieto dáta nevyžadovali žiaden rozsiahly automatizovaný a špecifický proces dolovania boli validované manuálne. Nižšie v tabuľke 5 ponúkame prehľad jednotlivých čŕt.

Tabuľka 5. Zoznam čŕt z kategórie historické dáta Bitcoinu

Atribút	Opis
open_price	Otváracia cena Bitcoinu na začiatku dňa
close_price	Zatváracia cena Bitcoinu na konci dňa
market_volume	Objem výmeny za posledných 24 hodín
market_cap	Počet Bitcoinov v obehu * cena
num_of_transaction	Počet transakcií za daný deň
miners_revenue	Počet vyťažených Bitcoinov * cena
output_volume	Suma všetkých transakčných výstupov za daný deň
total_trans_fees	Poplatky za 24 hodín

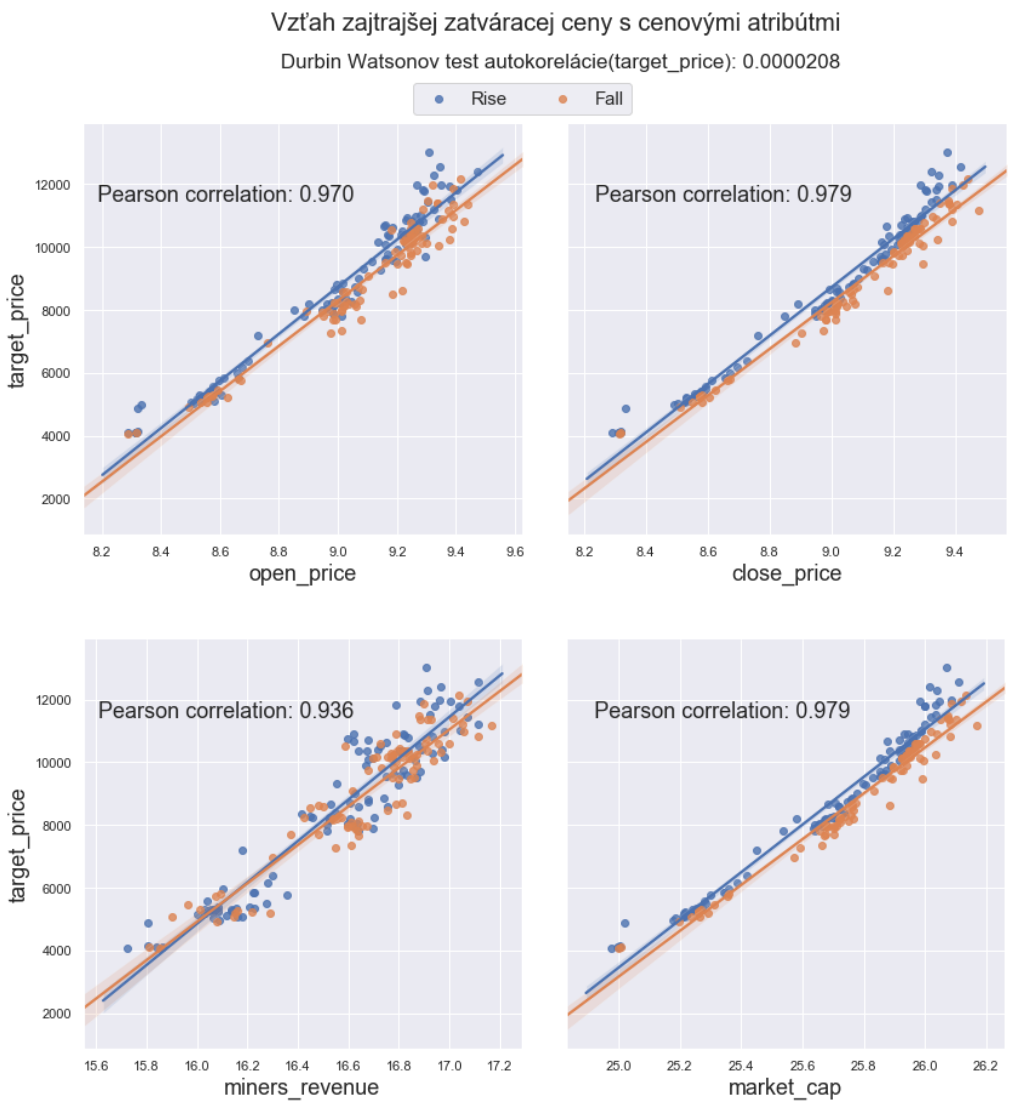
4.2.1. Exploratívna analýza – historické dáta Bitcoinu

V tejto kapitole prevedieme exploratívnu analýzu jednotlivých atribútov. Pozrieme sa na ich rozloženie a predpoklady pri určovaní a predikovaní ceny Bitcoinu. Z toho dôvodu bude väčšina atribútov porovnávaná so zatváracou cenou daného ale aj budúceho dňa, nakoľko práve táto cena má predpoklad byť naším cieľovým atribútom. Pri každom atribúte zvolíme najvhodnejší spôsob prezentácie a zhodnotíme jeho možné využitie pri predikcií. Na obrázku 6 nižšie môžeme porovnať vývoj jednotlivých atribútov v čase.



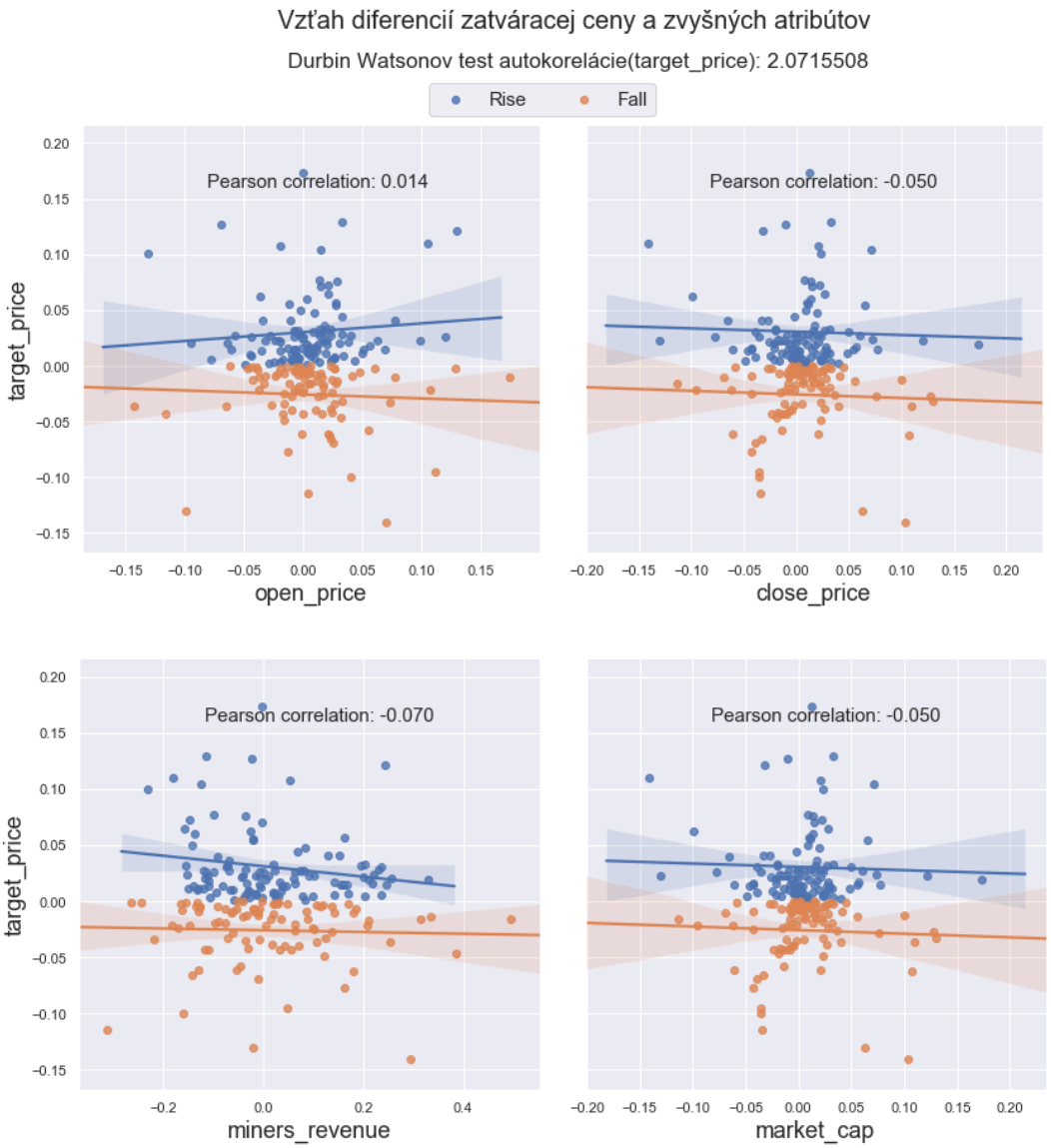
Obrázok 6. Vývoj čŕt historických dát v na časovej osi

Pre lepší a detailnejší pohľad na rozloženie daných atribútov porovnáme všetky hodnoty so zajtrajšou cenou Bitcoinu. Na nižšie zobrazenom obrázku 7 je možné pozorovať rozloženie atribútov so zajtrajšou cenou. Výber atribútov open_price, close_price, miners_revenue a market_cap bol zámerný, nakoľko jednotlivé hodnoty priamo predstavujú cenu Bitcoinu alebo sú z nej vypočítané. Tento fakt nezohľadňovali napríklad autori [8][16][17], ktorí aj tieto dáta z aktuálneho dňa zahrnuli do svojej dátovej množiny. To mohlo mať za následok umelé zvýšenie presnosti predpovedí nakoľko cieľový atribút je obsiahnutý v modifikovanej podobe v trénovacích dátach. Obzvlášť citlivý atribút môže byť market_cap, ktorý predstavuje cenu Bitcoinu prenásobenú objemom Bitcoinov v obehu. Práve počet Bitcoinov v obehu môžeme vo vzťahu k časovým radom považovať za konštantu. Tým sa tento atribút stáva použiteľný výhradne na úrovni historických dát a jeho použitie v deň predpovede je kontraproduktívne.



Obrázok 7. Zobrazenie vzťahu zajtrajšej ceny a porovnávaných atribútov

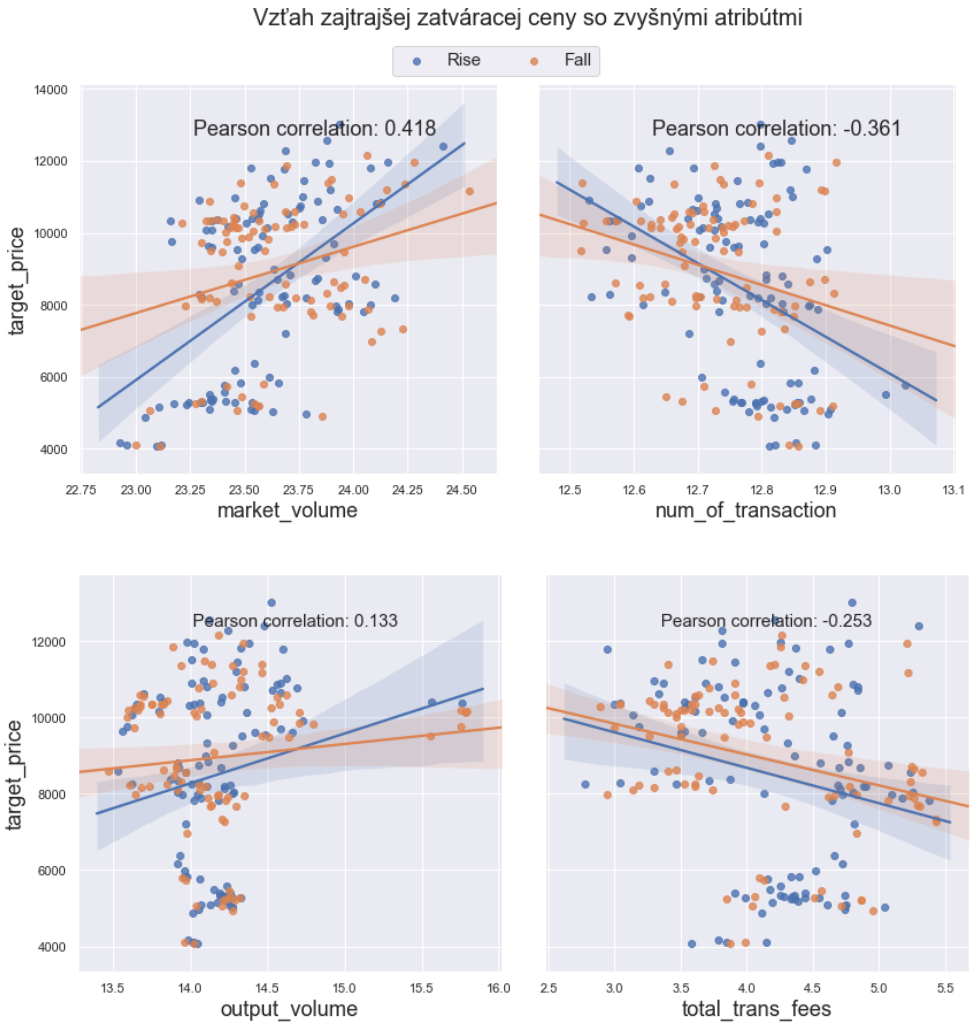
Ako môžeme vidieť jednotlivé atribúty v prvej štvorici výrazne korelujú aj vzhľadom na zajtrajšiu cenu. Dôvod je dopad vyššie spomínaného problému autokorelácie, ktorý Durbin Watsonov test s hodnotou 0,0000208 potvrdzuje. Nakoľko sledujeme atribúty vypočítavané práve z cien kryptomien jedná sa v princípe o autokoreláciu časového radu s časovým posunom t = 1. Z tohto dôvodu sme sa rozhodli aplikovať diferencie podľa vzťahu v rovnici 5. Voľne by sme mohli transformáciu opísať ako vzťah medzi percentuálnou zmenou ceny k zajtrajšiemu dňu s percentuálnymi zmenami atribútov v daný deň. Výsledok tohto posunu je možné vidieť na obrázku 8 nižšie kde vidíme výrazný prepad miery korelácie.



Obrázok 8. Rozloženie diferencie historických dát k zmene zajtrajšej ceny

Na druhej sade pozorovaní (obrázok 9) môžeme pozorovať zvyšné atribúty. Medzi najvýraznejšie ukazovatele patrí objem predaja na trhu, prípadne záporná korelácia počtu transakcií. Tento fakt je podporený aj obchodnými stratégiami, ktoré z logického hľadiska neobchodujú s menou, ktorá vykazuje nárast. Taktiež môžeme sledovať veľmi jemný súvis ceny a poplatkov za transakcie. Táto skutočnosť môže vplývať na bežné výmeny a transakcie na tomto trhu. Nakoľko ide o trh kde poplatok za transakciu sa pohybuje v desiatkach dolárov, je zrejmé, že transakcie s nižšou hodnotou a malou prioritou budú prevedené pri výhodnejších podmienkach. Vzhľadom na fakt, že ide o nezávislé časové rady a porovnávanie medzi nimi, autokorelácia nespôsobuje skreslenie pohľadu na jednotlivé dáta.

Z toho dôvodu dáta neprešli žiadnou transformáciou a je možné ich použiť v neupravenom tvare.



Obrázok 9. Vzťah atribútov nevypočítavaných z ceny k zajtrajšej cene

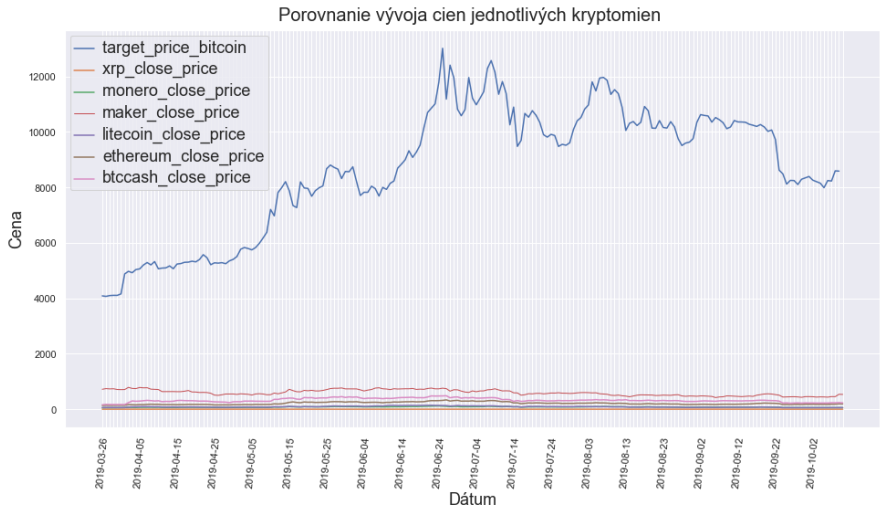
4.3.Dolovanie a analýza dát trhu kryptomien

Trh kryptomien predstavuje rovnaké portfólio pre každú jednu kryptomenu ako sme mohli vidieť v prípade historických dát Bitcoinu. Pre prvotnú analýzu sme preto vybrali 5 najzaujímavejších a najobchodovanejších kryptomien podľa [32], na ktoré sa zameriame pri našej analýze tejto skupiny dát. Dôvod sledovania tejto kategórie je zistiť previazanosť jednotlivých kryptomien navzájom a rýchlosť ich adaptovateľnosti vzniknutým situáciám na trhu. V prípade preukázania previazanosti medzi jednotlivými kryptomenami a odlišnými časovými úsekmi ich prispôsobivosti môže výrazne pomôcť pri určovaní budúcich hodnôt

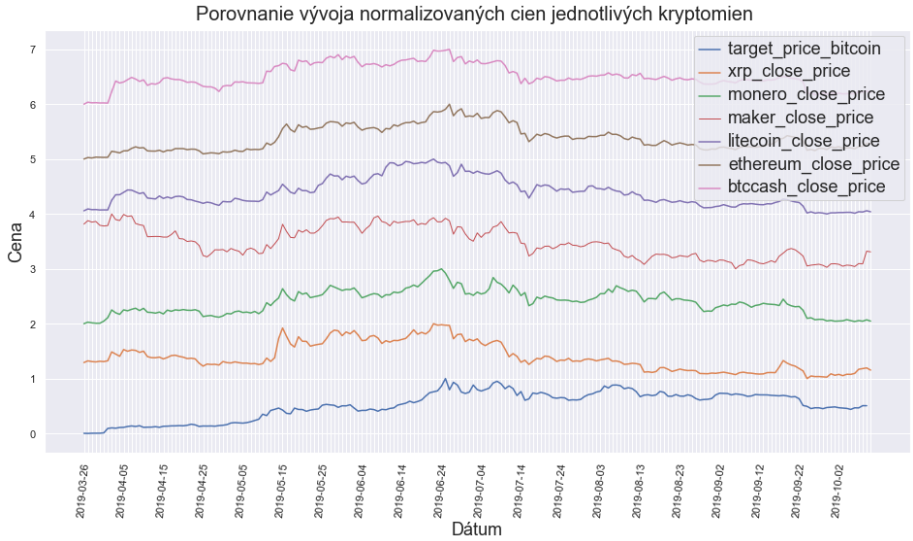
Bitcoinu. Nami vybrané a sledované kryptomeny sú BitcoinCash, Litecoin, Monero, XRP a Ethereum. Celý zoznam dolovaných čŕt je uvedený v prílohe E.

4.3.1. Exploratívna analýza - trh kryptomien

Z obrázku 10 uvedeného nižšie je jasné, že pohľad na surové dáta je ťažko porovnateľný z dôvodu vysokej ceny Bitcoinu oproti ostatným kryptomenám. Tento fakt môže prispievať k názoru, že práve Bitcoin môže byť pilierovou komoditou daného trhu. Z toho dôvodu sme sa rozhodli dáta normalizovať pre vhodnejšiu formu vizualizácie (obrázok 11). Napriek prvotnému nesúladu dát môžeme pozorovať výraznú podobnosť jednotlivých vývojov cien.

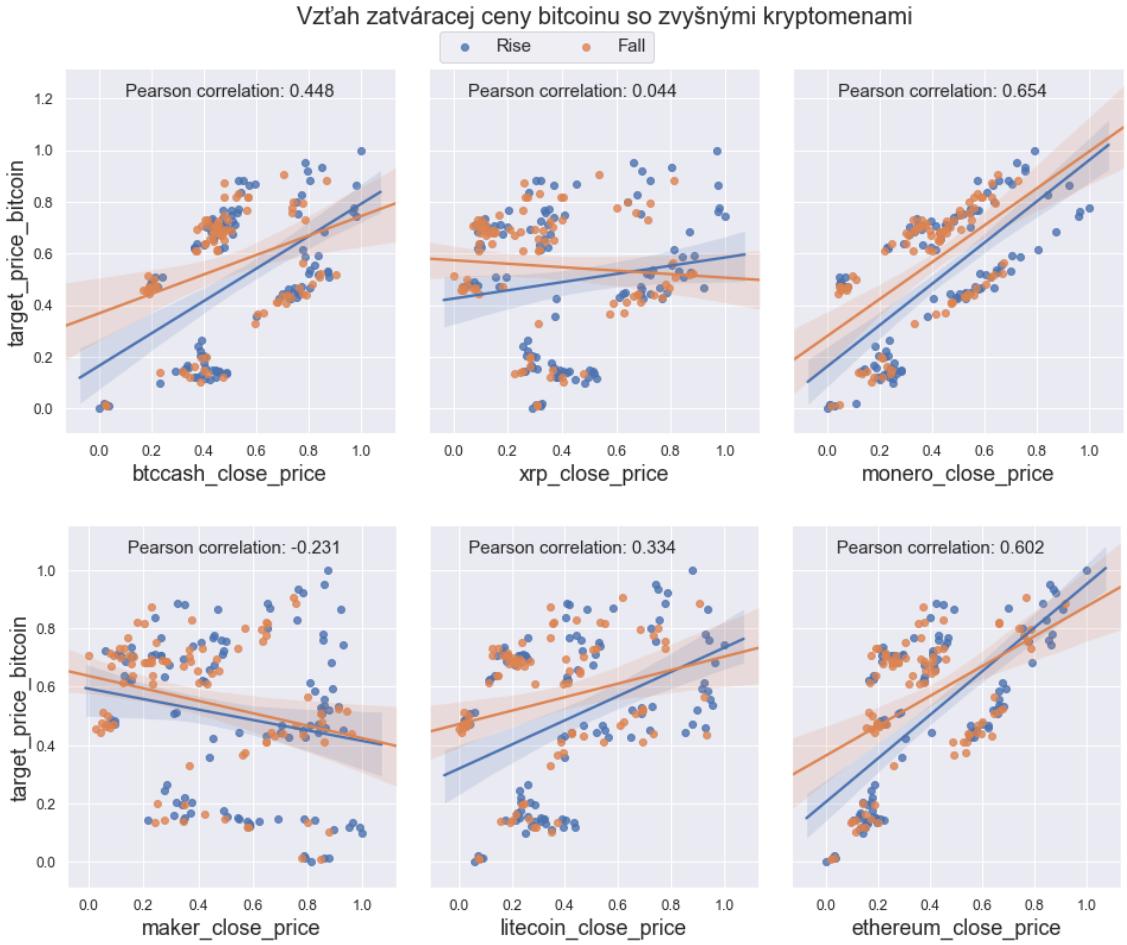


Obrázok 10. Vývoj cien jednotlivých kryptomien v čase (reálne hodnoty)



Obrázok 11. Vývoj cien jednotlivých kryptomien v čase (normalizované hodnoty)

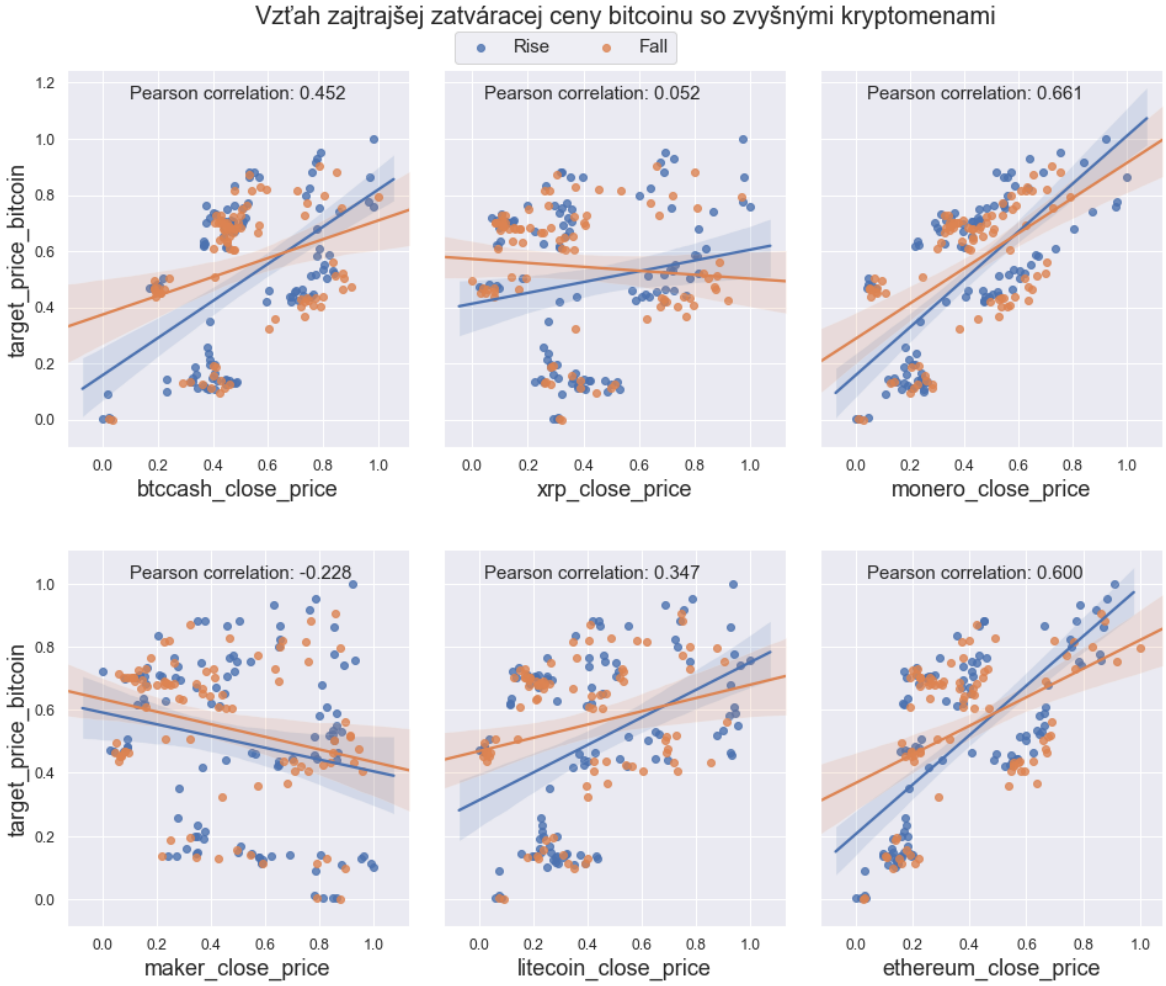
Táto skutočnosť sústredila našu pozornosť na bližšiu analýzu jednotlivých dát. K tomu nám pomohlo zobrazenie vzťahu ceny Bitcoinu k ostatným kryptomenám. Na obrázku 12 môžeme vidieť porovnanie jednotlivých kryptomien súčasného dňa s cenou Bitcoinu. Vidíme, že viaceré kryptomeny vykazujú nie zanedbateľnú mieru korelácie, čo je podporené aj Pearsnovým korelačným koeficientom. Taktiež sme sa rozhodli pozorovať zmenu ceny v daný deň z dôvodu odhalenia možných zhlukov.



Obrázok 12. Vzťah zatváracej ceny so zvyšnými kryptomenami

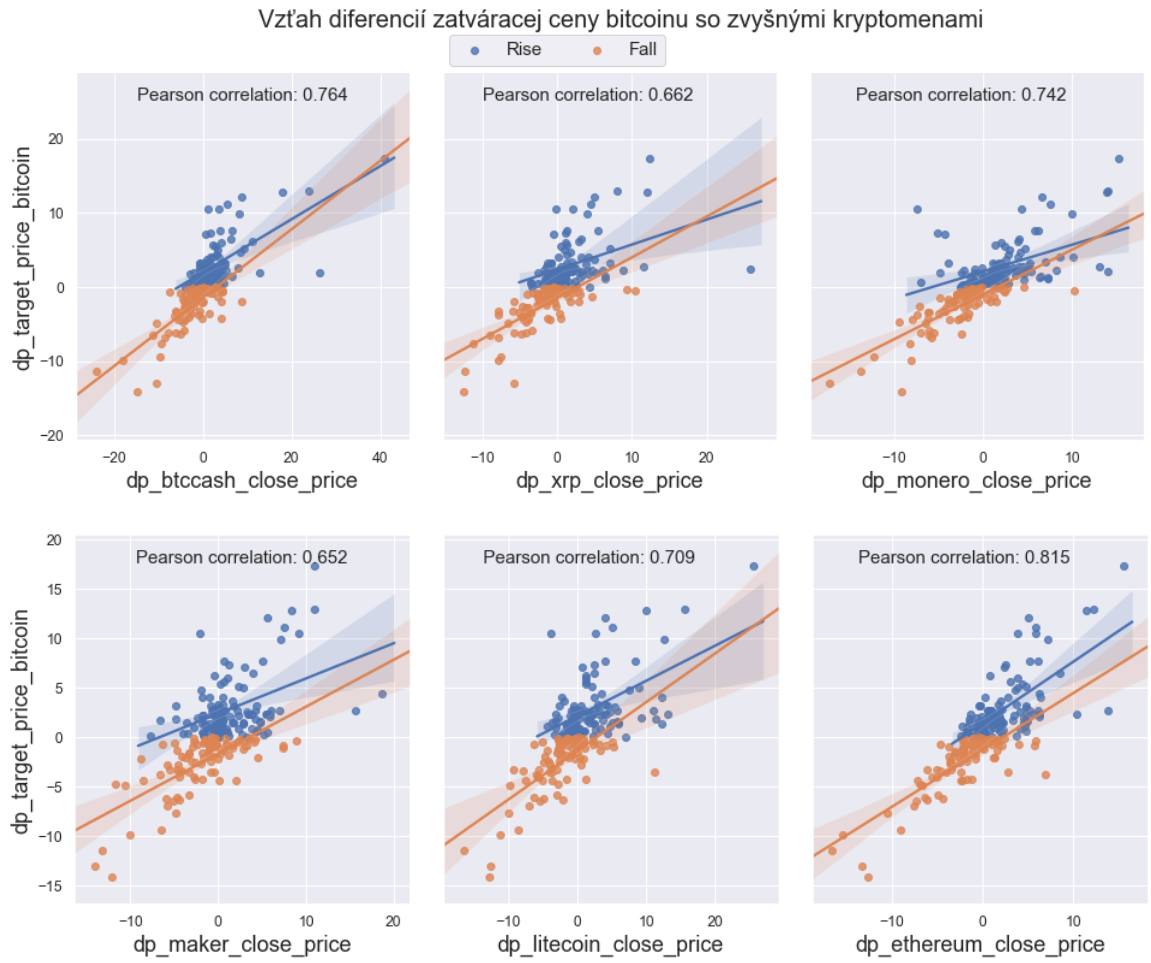
Z dôvodu vysokej korelácie jednotlivých cien sme sa rozhodli overiť koreláciu aj so zajtrajšou cenou Bitcoinu. Ako môžeme pozorovať (obrázok 13) výrazná zmena nenastala. Avšak ako sme spomenuli v kapitole 4.1.2 vyššie, môžeme stále sledovať parciálnu koreláciu a problém autokorelácie. Práve z tohto dôvodu sme sa rozhodli ceny jednotlivých kryptomien sledovať na úroveň ich diferencií. Diferencia bude pre tento prípad určená hodnotou zmeny k zajtrajšiemu dňu. Tento krok je znázornený na obrázku 14 a odhalil skutočnú previazanosť pohybov jednotlivých cien kryptomien. Jednotlivé zmeny daných kryptomien dokonca vykazujú vyšší koeficient korelácie ako samotné ceny, čo pre nás

predstavuje zaujímavý poznatok a dokazuje, že práve zmeny všetkých cien v daný deň sú výrazne korelujúce, pričom proces transformácie dát prebiehal v percentuálnej diferenciácií jednotlivých cien kryptomien k danému dňu.



Obrázok 13. Vzťah zatváracej ceny (zajtrajšej) so zvyšnými kryptomenami

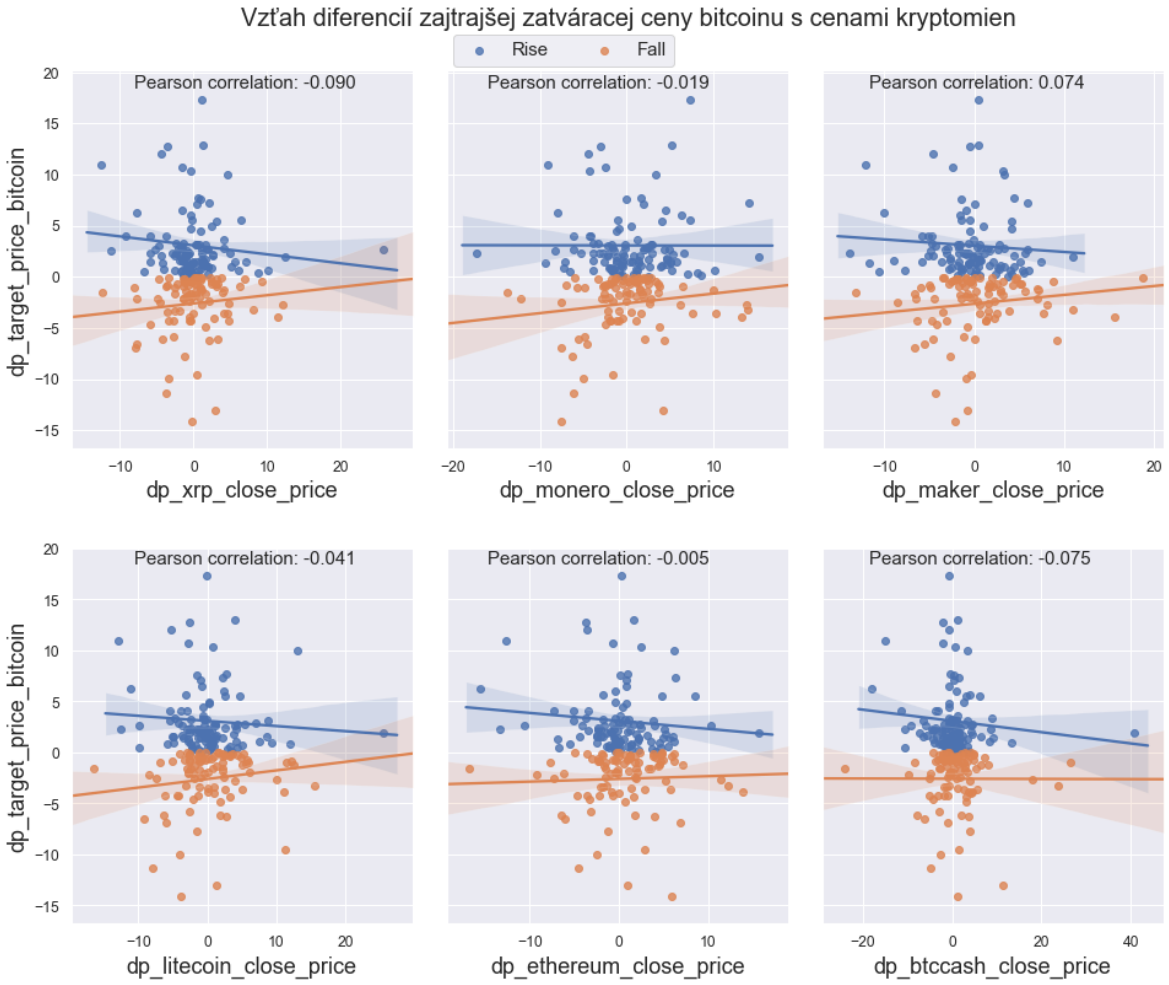
Toto pozorovanie ponúka zaujímavé zistenie o nárastoch a poklesoch ceny kryptomien v daný deň. Napríklad pri cene Etherea je možné pozorovať jav, kedy neexistuje pozorovanie s narastajúcou cenou Bitcoinu ak hodnota danej meny za daný deň klesla o viac ako 2,5%.



Obrázok 14. Vzťah diferencií zatváracej ceny (dnešnej) so zvyšnými kryptomenami

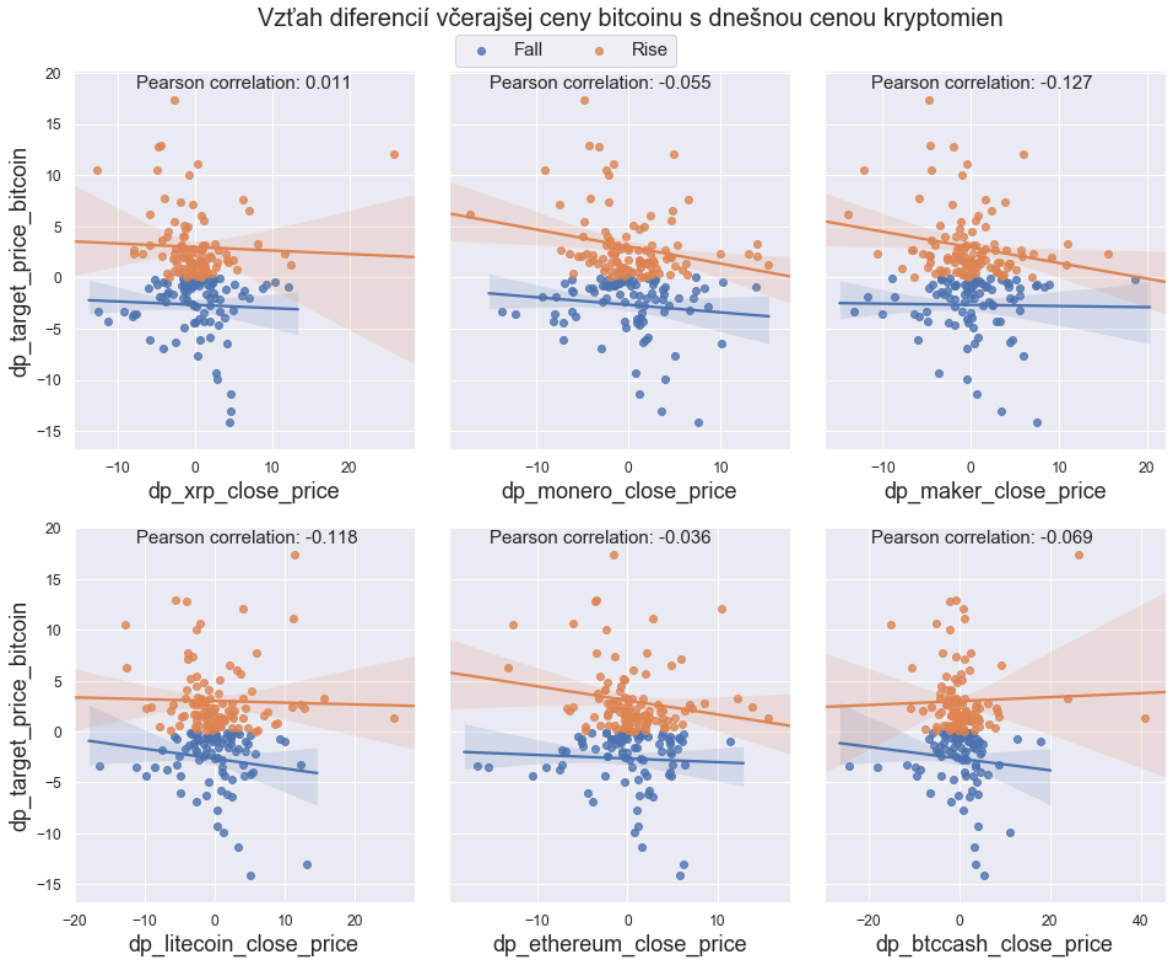
Následne sme analyzovali vzťah zajtrajšej zatváracej ceny a cien ostatných kryptomien. Na obrázku 15 vidíme výsledok tohto pozorovania. Ako môžeme vidieť korelácia jednotlivých cien sa úplne vytratila. Tento výsledok nám ukazuje, že neexistuje takmer žiadna lineárna závislosť medzi cenami zvyšných kryptomien so zajtrajšou cenou Bitcoinu. Dôvodov môže byť viacero. Ako sme v predchádzajúcom pozorovaní preukázali ceny kryptomien v daný deň výrazne korelujú aj na úrovni percentuálnych diferencií avšak zajtrajšia zatváracia cena, ktorá bola porovnávaná, je cena Bitcoinu o 24 hodín. Tento časový úsek môže byť priveľký a vplyv jednotlivých aspektov, ktoré ovplyvňujú cenu sa môže vplyvom času vytrácať. Ďalšou možnosťou môže byť opačná korelácia týchto dát, nakoľko ako sme naznačili v úvode nemáme znalosť o pilierovej komodite daného trhu. Faktom môže byť, že vývoj cien jednotlivých kryptomien nedokáže ovplyvniť zajtrajšiu cenu Bitcoinu ale naopak dnešná cena Bitcoinu môže mať vplyv na budúci vývoj ostatných kryptomien. Vzhľadom na túto hypotézu sme sa rozhodli overiť aj opačnú koreláciu, kde overíme vplyv percentuálnej zmeny ceny Bitcoinu na zmeny zajtrajších cien ostatných kryptomien. Taktiež treťou

nemenej pravdepodobnou možnosťou je reakcia cien na neznámy – nami nesledovaný podnet.



Obrázok 15. Vzťah diferencií zatváracej ceny (zajtrajšej) so zvyšnými kryptomenami

Nižšie (obrázok 16) uvádzame prípad porovnania opačnej korelácie a to vzťahom medzi včerajším pohybom ceny Bitcoinu a dnešným pohybom cien ostatných kryptomien. Cieľom pozorovania je preukázať mieru vplyvu vývoja dennej ceny Bitcoinu na ostatné kryptomeny. Ako je možné vidieť výsledok, je rovnaký ako pri predošlom skúmaní, čo môže znamenať, že celková korelácia je spôsobená na nižšej ako dňovej granularite alebo nesledujeme faktor ovplyvňujúci vývoj všetkých kryptomien vrátane Bitcoinu.



Obrázok 16. Vzťah diferencií zatváracej ceny (včerajšej) so zvyšnými kryptomenami

4.4. Dáta zo sociálnej siete Twitter

Dáta zo sociálnych sietí boli využité v mnohých prístupoch [4][11][13], autori však z dôvodu zložitého a časovo náročného prístupu k týmto dátam disponovali len krátkymi časovými úsekmi (menej ako 30 dní), ktoré Twitter ponúka na komerčnej báze. My sme zvolili formu verejne dostupnú, avšak dáta sme museli spracovávať v reálnom čase, čo si vyžiadalo implementáciu sledovacieho skriptu, ktorý bol nasadený na server s filtračnými nastaveniami na správy týkajúce sa Bitcoinu. Zahrnuté boli kľúčové slová (z angl. hashtag) ale aj samotné spomenutie slova v texte. Celý proces pozostával z nasledujúcich krokov:

1. Nastavenie API prístupu s potrebnými konfiguráciami
2. Filtrácia a poslanie dát v .json formáte zo strany Twittru
3. Extrakcia vybraných čŕt z poskytnutých dát
4. Vloženie záznamu ako riadok .csv súboru

4.4.1. Proces spracovania

Pri spracovávaní dát poskytnutých API službou sme museli riešiť nasledujúce problémy:

- veľkosť dát,
- rýchlosť spracovávania,
- uniformnosť dát.

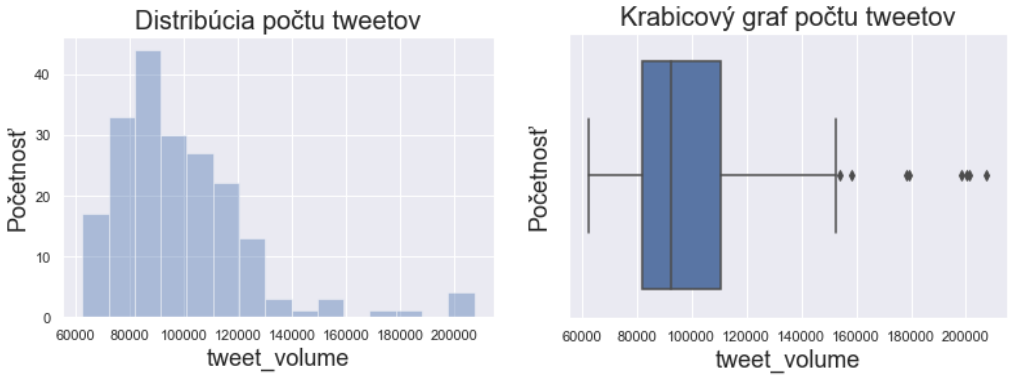
Prvým problémom bola veľkosť jednotlivých .json súborov. Pri prvotnom testovaní v závislosti od nastavenia účtu odosielateľa sa veľkosť súboru pohybovala od 5KB do 8KB čo s následne zisteným denným priemerom 60 000 - 80 000 predstavovalo v priemere 500MB dát za jeden deň. Z tohto dôvodu sme sa rozhodli json súbory spracovávať hneď na úrovni nasadeného skriptu. Analýzou poskytnutých atribútov sme vybrali sadu čŕt, ktorá bude extrahovaná už v procese získavania dát. Tieto črty sú:

- date – časová známka vytvorenia tweetu,
- id – identifikačný kľúč tweetu,
- user_id – identifikačný kľúč odosielateľa tweetu,
- user_follower_count – počet sledovateľov odosielateľa,
- user_favourites_count – počet stránok, ktoré má odosielateľ označené ako obľúbené,
- user_statuses_count – počet statusov daného odosielateľa,
- user_created_at – dátum založenia účtu odosielateľa,
- retweeted – pravdivostný príznak či sa jedná o preposlaný tweet,
- tweet_text – text v prípade krátkej správy,
- tweet_full_text – plný text v prípade správy, z ktorej bol vytvorený náhľad,

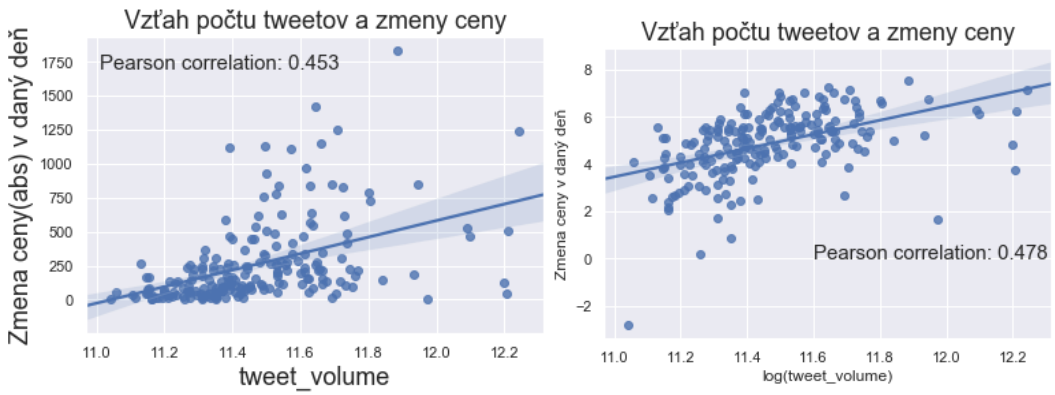
Tieto črty boli vybrané aj pre iné prípady použitia ako je ten náš, nakoľko jedným z cieľov práce je vybudovanie dátovej množiny. Vo vzorovom json súbore sú polia, ktorých názov evokuje vysokú použiteľnosť pre náš prípad použitia (počty páčikov, počet zdieľaní,..) avšak vzhľadom na živý tok dát sú tieto atribúty nulové a preto ani nie sú zahrnuté do nášho výberu finálnych atribútov. Týmto prístupom sme zredukovali veľkosť denných dát z pôvodných 500MB na 70MB. Tento fakt prispel aj k zlepšeniu rýchlosti spracovávania hneď na dvoch úrovniach. Prvá je proces spracovávania dát na úrovni API a druhá na úrovni následného spracovávania a extrakcie čŕt zo samotných tweetov (analýza sentimentu, čistenie, ..). Taktiež sa týmto prístupom dáta stali uniformnými pre uchovávanie v lepšom formáte pre budúce spracovanie ako je json. Výsledkom tohto procesu je sada denných .csv súborov, s vyššie spomenutými parametrami vo forme stĺpcov, kde riadky predstavujú jednotlivé tweety daného dňa.

4.4.2. Prvotná analýza

Po vytvorení prvej verzie dátovej množiny zo sociálnej siete Twitter sme pre potreby prvého experimentu analyzovali hodnoty z daného časového obdobia (26.03.2019 – 10.10.2019). Črta, ktorá bola prvá analyzovaná je počet tweetov za deň. Práve táto črta už počas samotného zberu spôsobovala výkyvy veľkostí jednotlivých denných .csv súborov. Nižšie na obrázku 17 je zobrazené rozloženie a krabicový graf práve pre túto črtu. Taktiež môžeme vidieť, že počet tweetov sa pohybuje od 62468 po 207522 pričom jeho medián je 92661 tweetov za deň. Ako môžeme pozorovať, niektoré hodnoty nadobúdajú extrémne hodnoty a spôsobujú skreslenie dát. Aj z toho dôvodu sme sa rozhodli zjemniť dopad vysokých hodnôt logaritmickou funkciou. Ich odstránenie neprichádza do úvahy, nakoľko práve tieto extrémne hodnoty môžu prezrádzať dianie na trhu. Na nižšie uvedenom obrázku 18 môžeme vidieť príklad vysokej heteroskedasticity, ktorú sme efektívnym spôsobom transformovali pri zachovaní výpovednej hodnoty daného pozorovania.



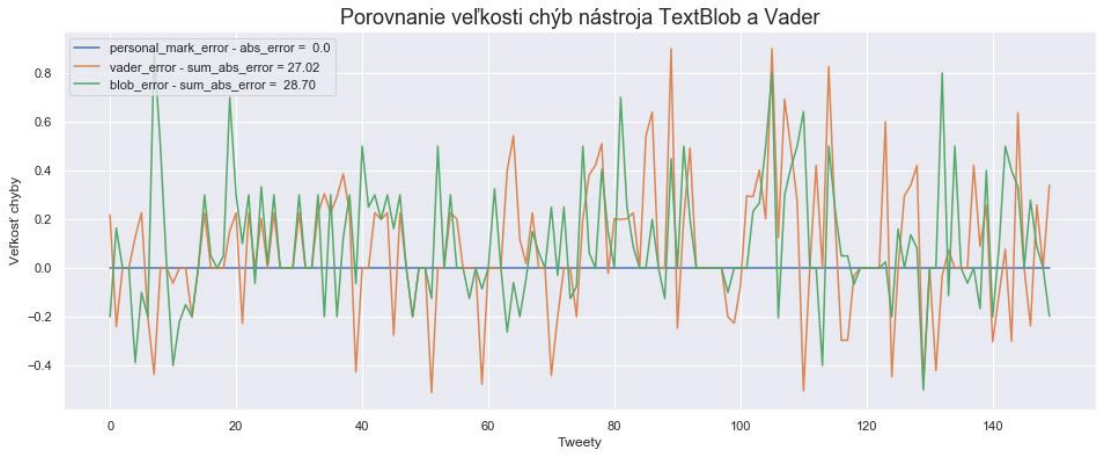
Obrázok 17. Zobrazenie početnosti správ zo sociálnej siete Twitter v nami sledovanom období



Obrázok 18. Zobrazenie korelácie denného počtu tweetow a absolútnej zmene ceny v daný deň

4.4.3. Extrakcia sentimentu z tweetov

V predchádzajúcej kapitole sme preukázali previazanosť denného počtu tweetov s cenovými výkyvmi trhu. Pomohli sme si však absolútnou hodnotou zmeny ceny, čo je v praxi vylúčené a musíme nájsť črtu, ktorá nám pomôže určiť smer výkyvu cenovej hladiny. V tomto ohľade nám môže byť nápomocný sentiment jednotlivých tweetov, ktorý môže poukazovať na náladu prispievateľov na sociálnej sieti Twitter a tým určiť smer vývoja ceny. Pri sociálnych sieťach, obzvlášť pri Twittri sa jedná o maximálne 280 znakové správy neštruktúrovaného charakteru. S ohľadom na tento fakt sme pri výbere uvažovali nad dvoma nástrojmi na extrakciu sentimentu, konkrétne Vader a TextBlob. Pre otestovanie nástrojov sme previedli ručné označenie sentimentu 150 tweetov následne sme rovnakú vzorku nechali analyzovať oboma nástrojmi a vypočítali absolútnu chybu každého z nástrojov. Výsledok tohto pozorovania je znázornený na obrázku 19.



Obrázok 19. Porovnanie chýb nástrojov TexBlob a Vader oproti nami určenej vzorke

Nakoľko nami vytvorená vzorka je triviálne malá, nevyberali sme exaktne podľa výsledku tohto testu. Zohľadnili sme napríklad aj formu výstupov jednotlivých nástrojov. TextBlob vracia ako výstup jedinú hodnotu v škále <-1, 1> ktorá predstavuje sentiment daného tweetu. Vader okrem toho, že je špecializovaný na použitie v prostredí sociálnych sietí (zohľadňovanie malých, veľkých písmen, emotikony,...) vracia sentiment v jednotlivých kategóriách a to:

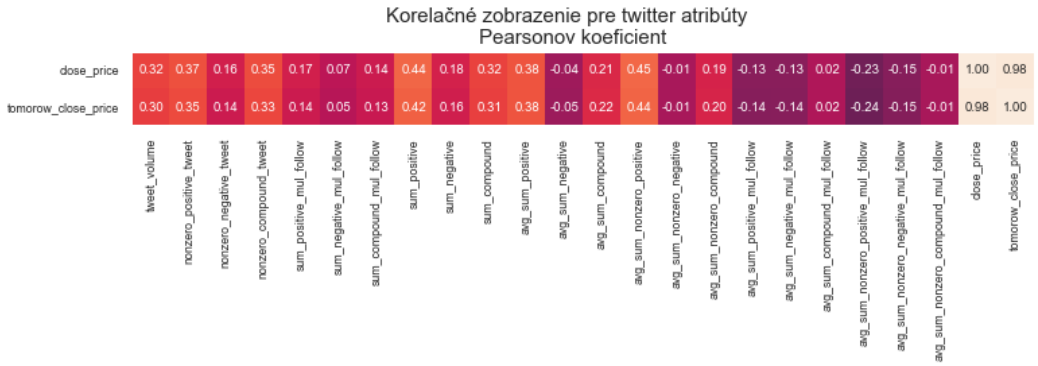
- positive – miera pozitivity daného tweetu z intervalu $< 0,1 >$,
- negative – miera negativity daného tweetu z intervalu $< 0,1 >$,
- neutral – miera neutrality daného tweetu z intervalu $< 0,1 >$,
- compound - agregovaná hodnota z troch predchádzajúcich z intervalu $< -1,1 >$.

Vzhľadom na ďalšiu použiteľnosť a nie výrazný rozdiel pri porovnávaní nástrojov sme vybrali Vader, s ktorým budeme ďalej pracovať. Tento krok nám vďaka rozloženiu celkového sentimentu ponúka možnosť sledovať určitú oblasť tweetov a výrazne zvyšuje použiteľnosť čŕt pre budúcu prácu.

Proces extrakcie sentimentu bol netriviálnou záležitosťou, nakoľko nami spracovávané obdobie (200 dní) obsahovalo viac ako 17GB textových dát. Samotný sentiment bol extrahovaný z 19 564 852 tweetov, čo spolu s ostatnými črtami predstavovalo viac ako 70 hodín čistého hardvérového času. Z tohto dôvodu sme metódu na extrakciu sformulovali tak, aby nám extrahovala čo najviac čŕt aj za cenu multikolienarity a nevyužiteľnosti v budúcom spracovávaní údajov. Nižšie uvádzame agregované hodnoty, pričom v prílohe C uvádzame zoznam všetkých dolovaných čŕt:

- tweet_volume – počet tweetov za daný deň
- nonzero_compound_tweet – počet nenulových sumárnych tweetov za daný deň
- sum_compound_mul_follow – Σ (sumárny sentiment * počet followerov)
- sum_compound – Σ (sumárny sentiment)
- avg_sum_compound – Σ (sumárny sentiment)/ tweet_volume
- avg_sum_nonzero_compound - Σ (sumárny sentiment)/ nonzero_compound_tweet
- avg_sum_compound_mul_follow - avg_sum_compound * počet sledovateľov
- avg_sum_nonzero_comp_mul_follow - avg_sum_nonzero_compound * sledovatelia

Vzhľadom na početnosť jednotlivých atribútov sme sa rozhodli previesť prvotnú analýzu pomocou korelačnej matice, ktorá bude sledovať ich lineárnu závislosť pomocou Pearsonovho korelačného koeficientu. Pre lepší prehľad sme túto maticu orezali (obrázok 20), nech porovnáva koreláciu voči dvom základným atribútom a to dnešnej zatváracej cene a zajtrajšej zatváracej cene.



Obrázok 20. Časť korelačnej matice pre ceny a jednotlivých čŕt extrahovaných z tweetov

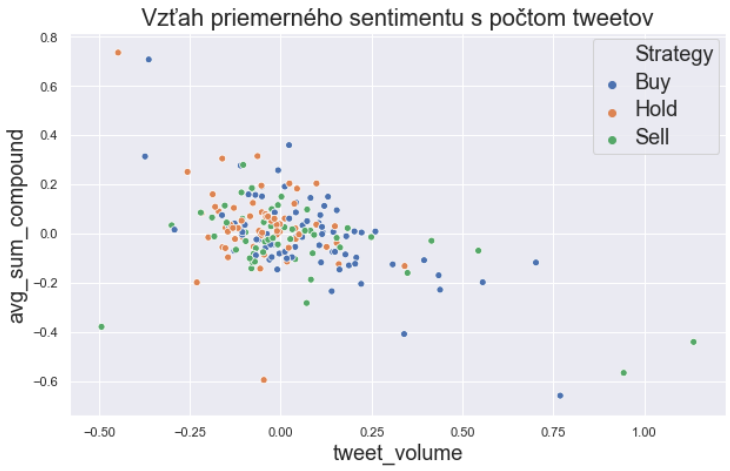
Z korelačného zobrazenia môžeme pozorovať viaceré korelácie. Avšak prihliadnuc na princíp časových radov môže byť čiastočne zavádzajúca. Korelačná matica v niektorých prípadoch porovnáva dvojice atribútov bez zohľadnenia časového aspektu. Vplyv a závažnosť absencie tejto informácie predvedieme na nasledujúcom príklade.

Uvažujme cieľovú cenu na úrovni 5000$ a atribút *sum_positive* v hladine 120 000. Je zrejmé, že z týchto dvoch údajov nevieme presne určiť ani smer ani mieru vývoja ceny. Celkovo môžu nastať až 4 scenáre s identickým výsledkom:

- hodnota sentimentu je výsledkom nárastu ceny na hodnotu 5000$,
- hodnota sentimentu je výsledkom poklesu ceny na hodnotu 5000$,
- cieľová cena je následkom nárastu miery sentimentu na hodnotu 120 000,
- cieľová cena je následkom nárastu miery sentimentu na hodnotu 120 000.

Otázkou však môže byť aj relevantnosť zohľadnenia minulých pozorovaní. Na základe uvedených scenárov je zrejmé, že nájsť vhodnú reprezentáciu, pre niektoré atribúty bude náročné. Taktiež forma transformácie u viacerých atribútov bude predstavovať netriviálnu záležitosť. Z tohto dôvodu kvôli jednoduchosti budeme uvažovať iba dva stavy ceny a to nárast a pokles. Takéto zobrazenie umožní lepšiu vizualizáciu jednotlivých vzťahov a pomôže jednoznačnejšie oddeliť závislé dáta. Taktiež vzhľadom na vysokú početnosť jednotlivých atribútov empiricky vyberieme základné, ktoré odprezentujeme nižšie.

Nižšie (obrázok 21) uvádzame výsledné rozloženie počtu tweetov a priemerného sentimentu vo vzťahu k cene. Všetky jednotlivé atribúty boli transformované na úroveň percentuálnych diferencií.



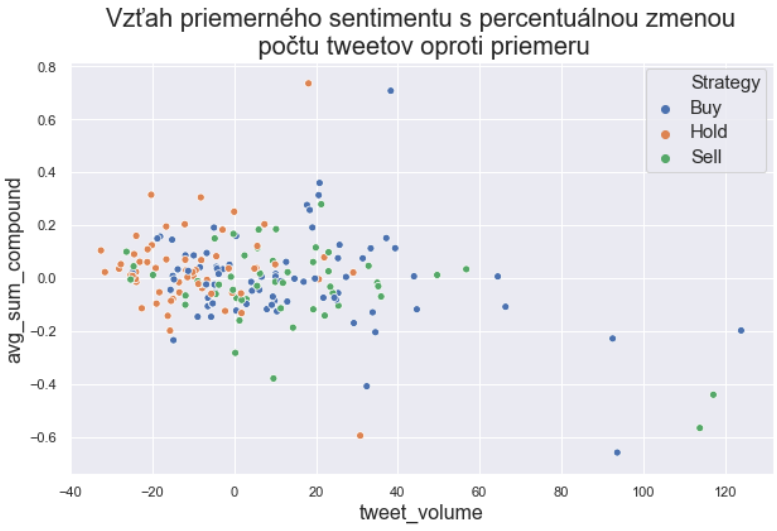
Obrázok 21. Rozloženie jednotlivých kategórií vzhľadom na priemerný sentiment a denný počet tweetov

Ako môžeme vidieť jednotlivé pohyby cien sa prekrývajú a nedá sa jednoznačne určiť zhluk nárastu alebo poklesu ceny vzhľadom na sledované atribúty. Napriek tomu v intervale objemu tweetov <-.50,0> je pozorovateľný vyšší výskyt poklesov ceny. Vzhľadom rozloženie údajov a koncentráciu sme sa rozhodli obmeniť stratégiu porovnávania. Pre lepšie oddelenie dát sme obmenili kategórie nárast/pokles tradičnou obchodnou stratégiou drž/predaj/kúp, kedy nakupujeme, keď má cena tendenciu rásť, predávame pri poklese a držíme pri nevýraznej zmene. Túto stratégiu sme aplikovali na úrovni +/- 1% zmeny ceny, čím sme vytvorili novú kategóriu. Na obrázku 22 nižšie môžeme vidieť predchádzajúce zobrazenie s ohľadom na novovzniknutú kategóriu. Ako môžeme pozorovať s poklesom objemu tweetov vzniká trend poklesu ceny prípadne minimálnou zmenou ceny. Rovnaké zobrazenie pre dané kategórie sme sa rozhodli previesť na viacerých atribútoch, ktoré boli intuitívne vybrané celej sady čŕt zo sociálnej siete Twitter.

Explicitnou analýzou objemu tweetov sme narazili na výrazne opakujúci sa jav, kedy pokles objemu tweetov nepredstavoval automaticky pokles ceny. Dôvodom je vysoká senzitivita tejto črty, nakoľko predstavuje chovanie sa používateľov Twittru v určitom okamihu. Tá ma za následok reakciu na výkyv ceny, ale jej návrat do normálu (pokles), skôr predstavuje stabilizáciu ceny, nie nutne jej pokles. Toto tvrdenie sme sa rozhodli overiť postupom kedy objem tweetov je transformovaný nasledujúcou formulou:

$$x = \left(\frac{denný\ počet\ tweetov}{mean(počet\ tweetov)} - 1\right) * 100$$

Rovnica 6. Vzťah pre výpočet relevantného pohybu tweetov



Obrázok 22. Rozloženie jednotlivých kategórií s využitím rovnice 6.

Aplikovaním tohto postupu môžeme vidieť na obrázku 22 jemnejšie rozdelenie jednotlivých kategórií. Aj keď sú jednotlivé kategórie premiešané a nevytvárajú sa jasne definované zhluky minimálne pre kategóriu Hold sa formuje silný vzor a väčšina jeho výskytov je zápornej časti pohybu objemu tweetov. Taktiež nemôžeme vylúčiť, že pridávanie dimenzií odhalí lepšie súvislosti medzi atribútmi zo sociálnych sietí. Aj tento stav a široká škála dolovaných atribútov navádza pre budúce využitie modelov s optimalizáciami pre automatický výber, nakoľko intuitívnym alebo pseudonáhodným výberom môžeme prísť o dôležité atribúty, ktoré nie priamo, ale parciálne vylepšujú celkový výsledok.

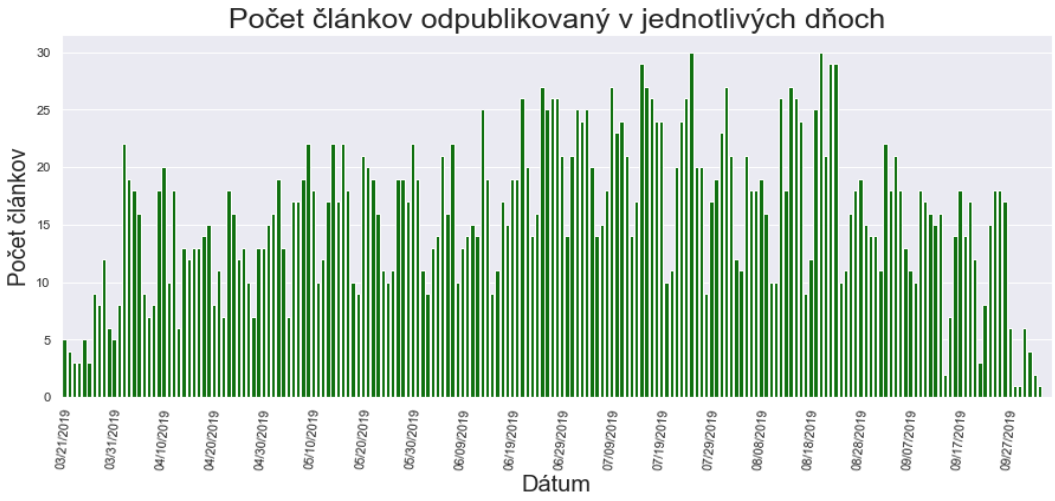
4.5. Dáta z internetových článkov

Ďalšou skupinou dát, ktoré sme zozbierali boli internetové články z rôznych portálov venujúcim sa problematike kryptomien. Špeciálne boli vyberané stránky, ktoré ponúkali štruktúrovaný obsah prípadne možnosť filtrácie, ktorej kryptomene je článok venovaný. Proces dolovania týchto dát bol obsiahly, nakoľko každý portál má vlastnú štruktúru stránok. Z toho dôvodu bolo nutné vytvoriť viacero rôznych dolovacích nástrojov. Analýzou jednotlivých článkov bola vytvorená uniformná šablóna dát, na základe ktorej sa extrahovali jednotlivé črty z týchto článkov. Zoznam týchto čŕt je znázornená v tabuľke 6 nižšie.

Tabuľka 6. Zoznam dolovaných atribútov z webových publikácií

Názov	Opis
source	Webová adresa portálu
url	Adresa ku konkrétnemu článku
author	Autor článku
date	Dátum publikovania
datetime	Časová známka publikovania
headertext	Titulok článku
subheaders	Titulky ďalších úrovní
contenttext	Obsah článku
tags	Tagy pripnuté k článku

Celkovo sa nám v nami zvolenom období 26.03.2019 – 10.10.2019 podarilo získať 3225 článkov, ktorých rozloženie v nami sledovanom období je zobrazené na obrázku 23 nižšie. Taktiež môžeme vidieť jednotlivé pravidelné výkyvy, ktoré predstavujú dni voľna (sobota, nedeľa).



Obrázok 23. Počet článkov publikovaný v jednotlivých dňoch nami sledovanými portálmi

4.5.1. Extrakcia sentimentu z článkov

Podobne ako pri tweetoch, aj pri článkoch sa stretávame s textovou formou dát a možnosťou extrakcie sentimentu v nich. V tomto prípade ide hlavne o nadpisy článkov a ich samotného obsahu. Napriek prvotnej podobnosti so správami zo sociálnej siete Twitter, proces extrakcie a analýza tejto sady dát vyžaduje rozdielny prístup. Oproti správam zo sociálnych sietí sa tu stretávame so štruktúrovaným a obsiahlejším textom. Články, ktoré máme k dispozícií sú písané výlučne v anglickom jazyku, vďaka čomu vieme využiť nástroje pre úpravu a čistenie textu bez výrazného obmedzenia. Proces extrakcie bol nasledovný:

- Odstránenie špeciálnych znakov a interpunkcie
- Odstránenie stop slov
- Extrakcia plnovýznamových slov
- Extrakcia sentimentu
- Výpočet agregovaných hodnôt

Odstránenie špeciálnych znakov a interpunkcie – v tomto kroku sme odstránili všetky znaky z textu netextového charakteru. Patria sem napríklad pomlčky, apostrofy, úvodzovky a pod.

Odstránenie stop slov - knižnica nltk [30] obsahuje kompletný slovník stop slov, ktorý sme aplikovali na všetky textové atribúty. Stop slová predstavujú spojky, neurčité členy a celkovo slová, ktoré nepridávajú z hľadiska spracovania textu žiadnu pridanú hodnotu.

Extrakcia plnovýznamových slov – týmto procesom sme všetky textové súčasti vyfiltrovali na úroveň plnovýznamových slovných druhov. Každé jedno slovo prešlo procesom

tokenizácie a na základe jeho POS (angl. part of speech) bolo zachované vo vete alebo bolo z nej vylúčené. Zamerali sme sa na extrakciu podstatných a prídavných mien, slovies a prísloviek.

Extrakcia sentimentu – v tomto momente sme pristúpili k procesu extrakcie sentimentu. Použitý bol rovnaký nástroj ako pri analýze sentimentu z tweetov. Sentiment bol extrahovaný na úrovni jednotlivých nadpisov článkov, ale i samotného obsahu daného článku. Taktiež za účelom porovnania efektivity čistenia textu sme sentiment extrahovali z čistených častí textu ale aj z nespracovaných.

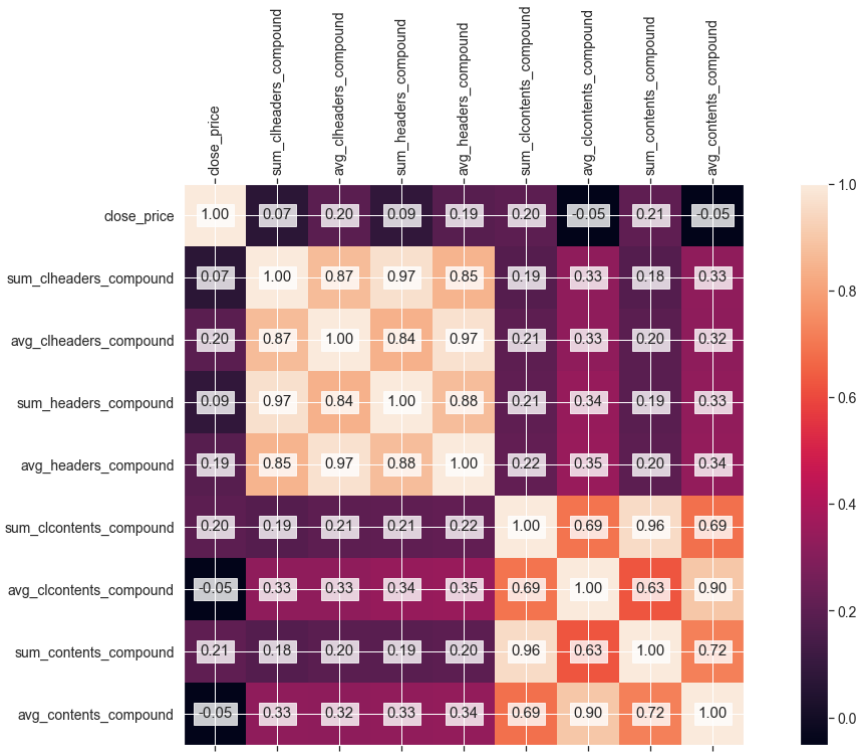
Výpočet agregovaných hodnôt – v tejto časti boli jednotlivé hodnoty agregované na úroveň jedného dňa tak, aby jednotlivé hodnoty boli reprezentatívne. Vo väčšine prípadov to znamená odvodenie priemerných hodnôt sentimentu pre dané dni.

Celkovo bolo extrahovaných niekoľko čŕt, ktoré budú následne využité pri hĺbkovej analýze tejto oblasti dát. Uvádzame základné denné črty pričom zoznam všetkých čŕt je uvedený v prílohe D:

- sum_clheaders_compound - suma hodnôt sentimentu zo spracovaných titulkov
- avg_clheaders_compound - priemerná hodnota sentimentu zo spracovaných titulkov
- sum_headers_compound - suma hodnôt sentimentu z nespracovaných titulkov
- avg_headers_compound - priemerná hodnota sentimentu z nespracovaných titulkov
- sum_clcontents_compound - suma hodnôt zo spracovaného obsahu
- avg_clcontents_compound – priemerná hodnota zo spracovaného obsahu
- sum_contents_compound – suma hodnôt z nespracovaného obsahu
- avg_contents_compound – priemerná hodnota z nespracovaného obsahu

4.5.2. Exploratívna analýza

Prístup, ktorý sme zvolili pri extrakcií sentimentu jednotlivých častí článkov a extrakciu na úrovni spracovaného a nespracovaného textu nám ponúka možnosť porovnať jednotlivé časti v korelačnej matici (obrázok 24). Táto matica nám pomôže vytvoriť prvotný pohľad na dáta a možnosť ich dodatočnej úpravy pre budúce použitie. Taktiež preukáže vplyv procesu čistenia na výsledné dáta a mieru zmeny týmto procesom dosiahnutým.



Obrázok 24. Korelačná matica pre atribúty z webových publikácií

Z korelačnej matice môžeme usúdiť, že neexistuje priama lineárna korelácia medzi cenou a sentimentov článkov publikovaných v daný deň. Tento záver však nie je prekvapujúci, nakoľko články oproti sociálnym sieťam nereagujú na zmeny trhu tak rýchlo ako sociálna sieť. Forma publikácie článku je v princípe výrazne odlišná ako uverejnenie svojho názoru na sociálnej sieti. Samotná príprava a tvorba článku trvá výrazne dlhšie ako napísanie 280 znakovej správy. Miera šírenia sa informácie medzi čitateľmi je výrazne pomalšia, nakoľko články sa musia rozšíriť medzi dané spektrum čitateľov a prípadný dopad jednotlivých článkov je časovo výrazne posunutý oproti komunikáciám v reálnom čase. Ďalším zaujímavým poznatkom je porovnanie čistených a nečistených textových častí. Tu vidíme, že jednotlivé sekcie sa nelíšia až tak výrazne. Máme však za dostatočne preukázané, že proces odstraňovania stop slov výrazne redukuje šum v dátach a tým pádom by mala byť práve táto hodnota presnejšia. Nemenej dôležitým poznatkom je fakt, že samotné nadpisy článkov sú len parciálne korelujúce so svojim obsahom. Toto pozorovanie v princípe dokazuje využitie známej praktiky, kedy sú titulky článkov písané zavádzajúco a zámerne skreslené pre prilákanie čitateľa na danú doménu (angl. clickbait).

4.6. Dáta z trhu akcií

Táto sekcia dát je obsiahnutá za účelom dokázania previazanosti trhu kryptomien s inými trhovými segmentami. Nakoľko Bitcoin je založený na technológií blockchain, pri ktorej je nemalý súvis s hardvérom, ktorý je nevyhnutný či už pre ťažbu alebo hľadanie najmenšej heš hodnoty. Vzhľadom na tento fakt sme sa rozhodli zahrnúť do pozorovania aj trh akcií kde sme obsiahli ceny akcií jednotlivých spoločností v IT sektore, napríklad Nvidia, Intel či AMD. Zahrnuli sme i parciálne súvisiace komodity ako ropu, ktorá je pilierovým elementom pre mnohé iné komodity, či kobalt, ktorý je kritickou surovinou pri výrobe grafických kariet.

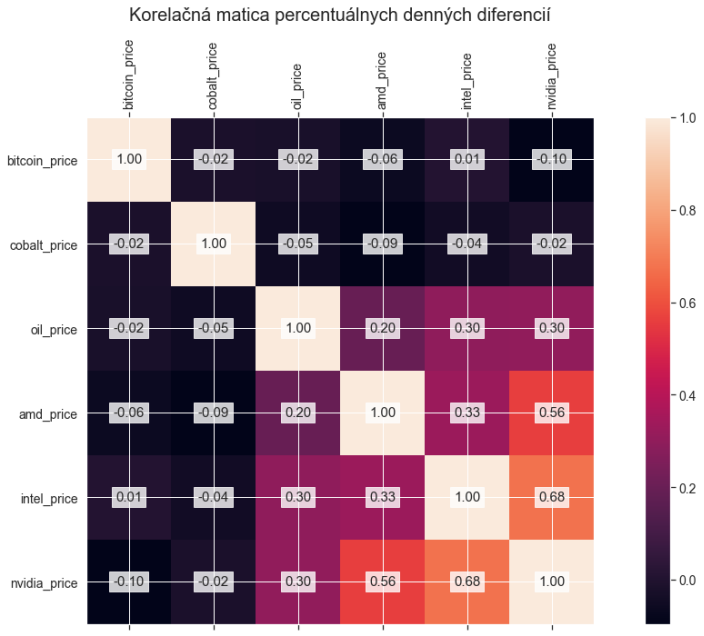
4.6.1. Exploratívna analýza

Exploratívna analýza bude v tomto segmente prevedená na dvoch úrovniach. Na prvej úrovni sa pozrieme na koreláciu celkových cien jednotlivých komodít navzájom prostredníctvom korelačnej matice. V druhom kroku sa pozrieme na korelácie denných diferencií čím eliminujeme vplyv autokorelácie a sústredíme sa špeciálne na percentuálnu diferenciu na úrovni dní. V tomto pozorovaní budeme brať v úvahu iba ceny vyššie spomenutých komodít. Avšak okrem týchto atribútov naša dátová množina disponuje aj dodatočnými črtami a ich kompletný zoznam je opäť možné vidieť v prílohe F.

Obrázok 25. Korelačná matica pre atribúty trhu akcií

Ako môžeme vidieť (obrázok 25) miera korelácie je medzi viacerými komoditami vysoká, avšak dodatočnou ručnou analýzou dát sme vyhodnotili nevhodnosť využitia ceny kobaltu, z dôvodu veľmi riedkych dát. Celkovo však táto matica poukazuje iba na podobnosť kriviek.

Ako sme spomenuli v úvode, väčšiu výpovednú hodnotu môže mať podobnosť na úrovni percentuálnych diferencií jednotlivých dní. Nanešťastie ani táto vizualizácia nemusí odhaliť koreláciu nakoľko analyzujeme podobnosť jednotlivých dní.



Obrázok 26. Korelačná matica pre atribúty trhu akcií (diferencie cien)

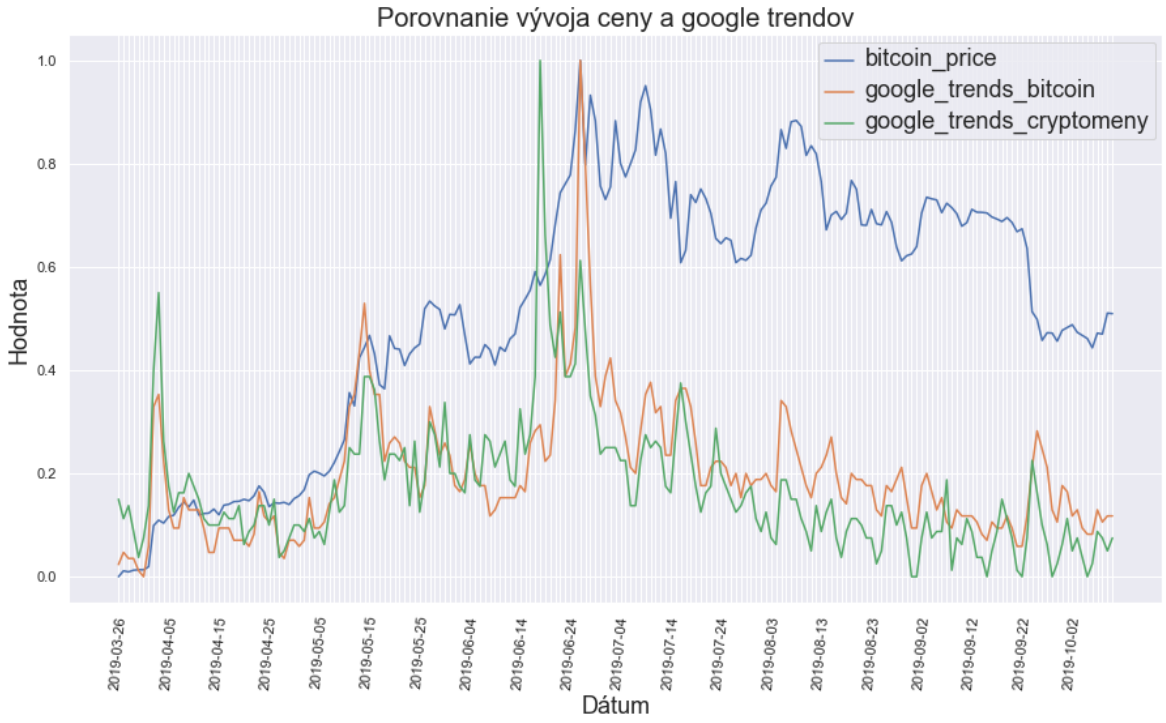
Výsledok pozorovania (obrázok 26) spočíva vo vytratení korelácie Bitcoinu s ostatnými komoditami. Niektoré korelácie boli zachované, čím sa preukázala ich skutočná previazanosť na úrovni sledovaných dní.

4.7. Dáta z Google Trends

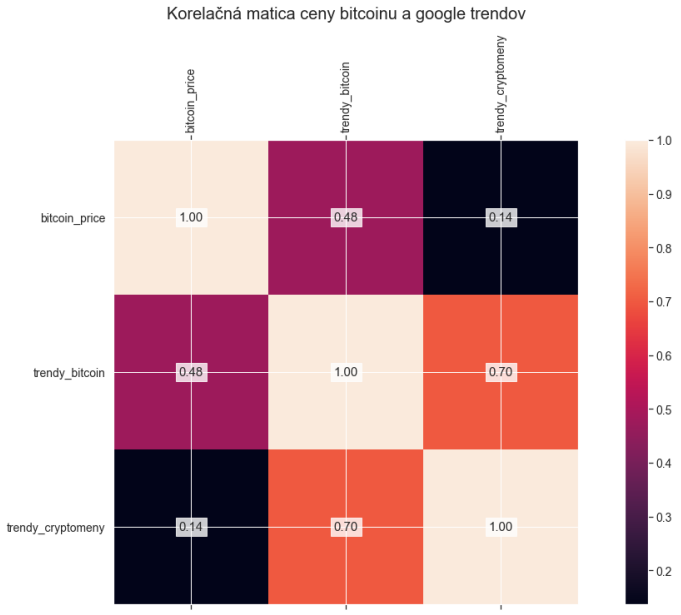
Táto časť priamo súvisí so všeobecnou mienkou a náladou spoločnosti. Google Trends predstavujú mieru vyhľadávania určitých fráz na dennej báze. Táto skutočnosť tak ako aj tweety, môže odzrkadľovať aktuálnu mieru záujmu širokého spektra populácie o Bitcoin. Dáta boli získané z verejne dostupnej služby spoločnosti Google, kde sme sa zamerali na vyhľadávanie fráz v dvoch kategóriách. Frázy obsahujúce slovo „Bitcoin“ a frázy všeobecnejšie obsahujúce slovo „Cryptocurrency“ pre širšie pokrytie.

4.7.1. Exploratívna analýza

Dáta z tohto zdroja predstavujú dennú mieru záujmu vyhľadávania vyjadrenú na škále 0-100 v danom sledovanom období. Nižšie na obrázku 27 uvádzame porovnanie vývoja týchto hodnôt normalizovaných na rovnakú mierku 0-1.



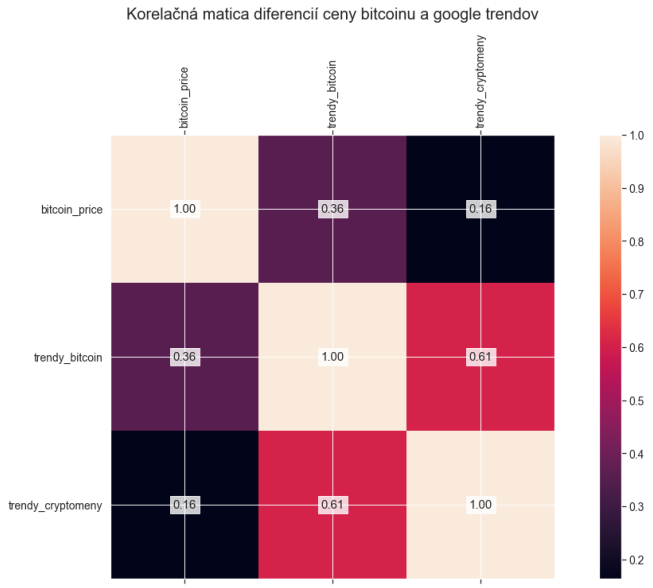
Obrázok 27. Vývoj miery vyhľadávania kľúčových slov s cenou Bitcoinu



Obrázok 28. Korelačná matica pre atribúty Google Trends s cenou Bitcoinu

Pre lepší pohľad na tieto hodnoty uvádzame aj korelačnú maticu uvedenej na obrázku 28. Google Trends pre vyhľadávanie Bitcoinu sú výrazne vyššie ako pre všeobecnejší názov.

Taktiež ako v predchádzajúcich kapitolách sa pozrieme na koreláciu diferencií. Táto závislosť je znázornená nižšie (obrázok 29), kde vidíme, že jednotlivé korelácie sa zjemnili.



Obrázok 29. Korelačná matica pre atribúty Google Trends s cenou Bitcoinu (diferencie)

5. Experimenty

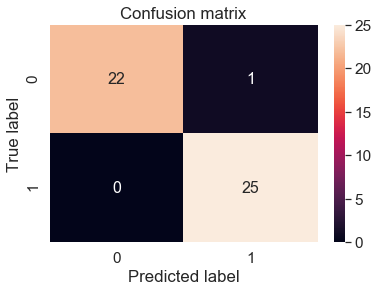
V tejto kapitole prevedieme niekoľko experimentov, na základe ktorých zhodnotíme poznatky z doterajšej analýzy. Jednotlivé experimenty budú poskytovať proces prípravy dát ale aj výber, konštrukciu, prípadne optimalizáciu zvolených metód. Každý experiment bude obsahovať vyhodnotenie úspešnosti zvolených metód.

5.1. Predbežný experiment – rekonštrukcia práce [17]

Ako sme v kontexte práce viackrát uviedli, cieľom je predikcia budúcej hodnoty Bitcoinu. Analýzou jednotlivých prác sme narazili na nekonzistenciu v určení cieľového atribútu. Napríklad v práci autorov [8][11][17] je cieľový atribút určený zatváracou cenou daného dňa. Tento prístup sa môže zdať ako správny, avšak treba si uvedomiť vzťah dát, ktoré využívame pri trénovaní. Autori v práci [17], dosahujú vysokú úspešnosť s použitím základných historických dát Bitcoinu a blockchainu. Nakoľko práca nie je oficiálna vedecká publikácia, ale ide o projekt členov Standforskej univerzity, ktorý bol spomenutý vo viacerých relevantných zdrojoch, rozhodli sme sa postup autorov analyzovať a zistiť príčinu vysokej úspešnosti. Autori využívajú sadu 16 čŕt pričom niektoré z nich pokrýva aj naša dátová množina. Ako cieľový atribút pre nich slúži zmena priemernej ceny daného dňa. To znamená, že svoju úlohu transformovali na binárny klasifikačný problém. Nakoľko nami vytvorená dátová množina nedisponuje explicitne priemernou dennou cenou ale otváracou a zatváracou cenou, tento atribút jednoducho vypočítame priemerom týchto dvoch hodnôt. Taktiež nám chýbajú niektoré atribúty, ktoré autori využili pri svojom riešení. Z toho dôvodu využijeme iba atribúty, ktorými disponujeme a vynecháme nami získané dodatočné črty. Nižšie uvádzame zoznam atribútov spoločných pre obe dátové množiny:

- market capitalization,
- miners revenue,
- number of TXN,
- TXN fees total,
- estimated transaction volume,
- differencie predošlých čŕt.

Autori deklarujú úspešnosť 95% pri využití modelu náhodného lesa. Nakoľko neuvádzajú špecifické nastavenie parametrov ponechávame základné nastavenia modelu z knižnice scikit. Nižšie môžeme vidieť maticu zámen (obrázok 30) výsledkov nášho modelu. Pričom model dosiahol celkovú úspešnosť 97.9%.



Obrázok 30. Matica zámen po replikovaní experimentu

Zhodnotenie prístupu - Dodatočnou analýzou vplyvu čŕt sme zistili, že najsignifikantnejšia črta je diferencia atribútu market_cap. Táto črta je vypočítavaná nasledujúcim vzťahom:

$$marketCap = počet\ Bitcoinov\ v\ obehu\ *\ cena\ Bitcoinu$$

Rovnica 7. Vzťah pre výpočet atribútu market_cap v praxi

Ak zohľadníme fakt, že počet Bitcoinov je relatívne pomaly narastajúca hodnota, využitie tejto črty je kontraproduktívne pri predikovaní cieľového atribútu. Opomenúc tento fakt, zmena priemernej ceny v daný deň v kontexte binárnej klasifikácie je v našom prípade nevhodná. Nami dolované atribúty, napríklad sentiment tweetov v daný deň môže byť priamou reakciou na zmenu ceny daného dňa, čím by náš model neriešil úlohu predikcie ale schopnosti opisu reakcie čŕt na vývoj ceny.

4.8. Predbežný experiment – denná predikcia

V tomto experimente sa zameriame na optimalizáciu predošlého riešenie s odstránením nedostatku – zle určeného cieľového atribútu. Z toho dôvodu bude naše riešenie v prvotnej fáze taktiež založené na binárnej klasifikácií. Tento prístup sme zvolili z dôvodu ľahkej porovnateľnosti dosiahnutých výsledkov. Úlohou bude predikcia vývoja ceny. Vzhľadom na predchádzajúci príklad nevhodne vybraného cieľového atribútu k použitým dátam, definujeme náš cieľový atribút nasledujúcim vzťahom :

$$zmena\ ceny_t = Cena\ Bitcoinu_{t+1} - Cena\ Bitcoinu_t$$

Rovnica 8. Vzťah pre výpočet nami určeného cieľového atribútu

Týmto vzťahom zabezpečíme, že žiaden atribút použitý pri trénovaní nebude reflektovať vývoj ceny daného dňa, nakoľko predikujeme zmenu ceny k zajtrajšiemu dňu. Dôležité je brať v úvahu aj samotnú cenu, nakoľko disponujeme otváracou aj zatváracou cenou daného dňa. Aby sme predišli problému uvedenému vyššie, prichádza do úvahy iba zatváracia cena Bitcoinu.

4.8.1. Návrh metód

Z dôvodu vhodnej komparácie modelov a skúseností iných autorov naše riešenie bude pozostávať z využitia nasledujúcich metód:

- náhodný les (RF),
- naivný Bayesov klasifikátor (GNN),
- podporný vektorový klasifikátor (SVC),
- k najbližších susedov (KNN).

Každý z modelov disponuje možnosťou odhadu istoty svojej predpovede na základe ktorej sa dajú modely efektívnejšie porovnávať. Práve táto možnosť bude pre nás podstatná, nakoľko jednotlivé modely budú medzi sebou porovnávané a analyzované.

4.8.2. Opis úpravy a výberu čŕt

Vzhľadom na množstvo čŕt získaných v procese dolovania a analýzy, musíme riešiť ich relevanciu a rozloženie. V časti exploratívnej analýzy jednotlivých kategórií sme poukázali na skutočnosť, že daný problém nemôžeme považovať za triviálny. V procese experimentu sa nám osvedčila metóda StandardScaler [31], ktorá je odporúčaná pre preškálovanie jednotlivých čŕt, nakoľko viaceré modely vyžadujú rozloženie dát v okolí 0. Taktiež jednotlivé modely disponujú atribútmi, ktoré umožňujú ovplyvniť proces učenia a môžu pomôcť pri výbere najvhodnejších čŕt. Pri RF budú primárne sledované parametre max_depth a min_samples_split. Naivný Bayesov klasifikátor nedisponuje parametrami, ktoré by mohli pomôcť pri selekcií vhodných čŕt. SVC obsahuje regularizáciu L1 (lasso), ktorá dokáže optimalizovať výber čŕt pri ich vysokom počte. KNN obsahuje parameter počtu susedov, ku ktorým sa má cieľová hodnota najviac priblížiť.

4.8.3. Výsledok Experimentu – binárna klasifikácia

Nakoľko nami zvolená množina disponuje celkovo 250 pozorovaniami je vhodné použiť krížovú validáciu. Konkrétne sme sa rozhodli pre K-1 krížovú validáciu. Tento prístup bol zvolený pre maximálne využitie nízkeho počtu pozorovaní pri procese trénovania. Pre zistenie dôležitosti jednotlivých kategórií v procese predikcie sme proces trénovania a testovania realizovali nad každou kategóriou separátne. Tento prístup nám pomôže lepšie zhodnotiť vplyv jednotlivých kategórií na celkový výsledok. V tabuľkách nižšie uvádzame jednotlivé výsledky modelov pri vybraných kategóriách čŕt.

Tabuľka 7. Úspešnosť jednotlivých modelov v definovaných kategóriách

Uspešnosť (Accuracy)					
	Historické dáta Bitcoinu	Twitter dáta, Google Trends	Webové publikácie	Trh akcií	Trh kryptomien
RF	0.58620	0.55172	0.53448	0.56896	0.60344
GNN	0.56896	0.50000	0.50000	0.55172	0.55172
SVC	0.53448	0.53448	0.48275	0.60344	0.51724
KNN	0.53448	0.50000	0.51724	0.53448	0.53448

Tabuľka 8. Presnosť a pokrytie jednotlivých modelov v definovaných kategóriách pre triedu nárast

Presnosť (Precission)/Pokrytie (Recall) pre predikciu nárastu - 1					
	Historické dáta Bitcoinu	Twitter dáta, Google Trends	Webové publikácie	Trh akcií	Trh kryptomien
RF	0.56 / 0.79	0.57 / 0.65	0.61 / 0.56	0.55 / 0.64	0.61 / 0.67
GNN	0.63 / 0.40	0.54 / 0.76	0.62 / 0.43	0.52 / 0.59	0.68 / 0.39
SVC	0.60 / 0.47	0.62 / 0.57	0.53 / 0.66	0.76 / 0.53	0.68 / 0.42
KNN	0.56 / 0.65	0.47 0.56	0.48 / 0.56	0.57 / 0.55	0.54 / 0.67

Tabuľka 9. Presnosť a pokrytie jednotlivých modelov v definovaných kategóriách pre triedu pokles

Presnosť (Precission)/Pokrytie (Recall) pre predikciu poklesu – 0					
	Historické dáta Bitcoinu	Twitter dáta, Google Trends	Webové publikácie	Trh akcií	Trh kryptomien
RF	0.65 / 0.38	0.52 / 0.44	0.44 / 0.50	0.60 / 0.50	0.60 / 0.54
GNN	0.54 / 0.75	0.33 / 0.16	0.41 / 0.61	0.59 / 0.52	0.49 / 0.76
SVC	0.48 / 0.62	0.42 / 0.48	0.39 / 0.27	0.48 / 0.73	0.42 / 0.68
KNN	0.50 / 0.41	0.54 / 0.45	0.56 / 0.48	0.50 / 0.52	0.52 / 0.39

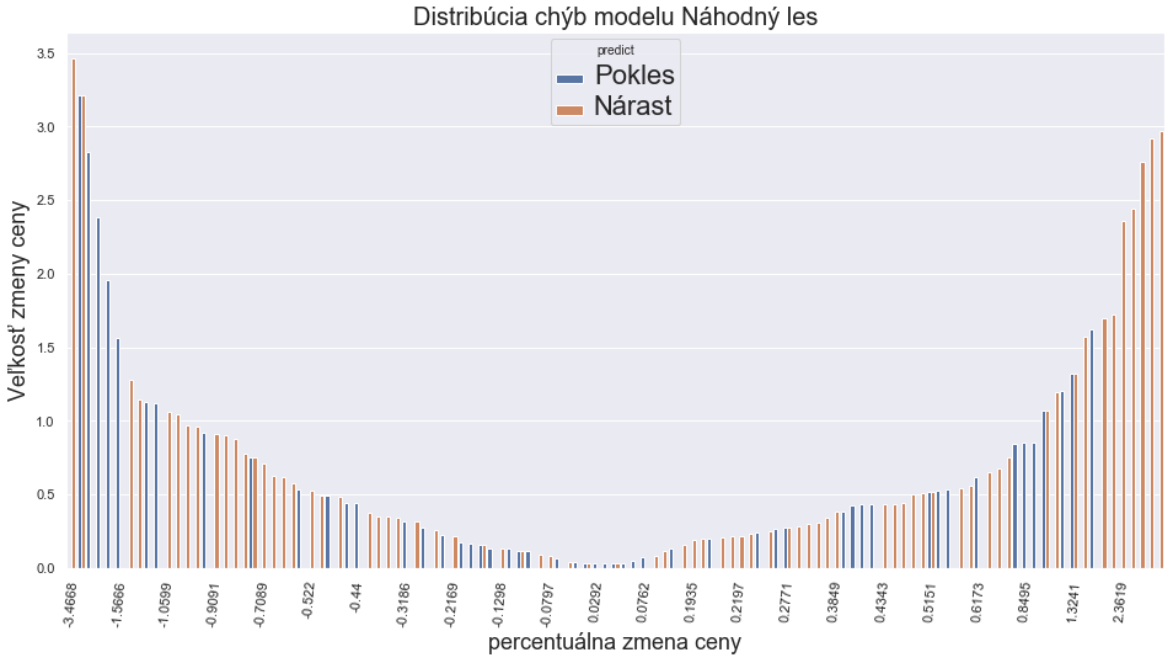
Tabuľka 10. Metrika AUC jednotlivých modelov pri použití špecifických kategórií

AUC jednotlivých modelov					
	Historické dáta Bitcoinu	Twitter dáta, Google Trends	Webové publikácie	Trh akcií	Trh kryptomien
RF	0.61356	0.63500	0.571691	0.63392	0.598214
GNN	0.66071	0.58909	0.549060	0.65113	0.572121
SVC	0.55889	0.55900	0.496394	0.66792	0.574494
KNN	0.59617	0.48506	0.550776	0.52270	0.436904

Tabuľka 11. Metriky modelov pri využití čŕt z celej dátovej množiny

Výsledky na celej dátovej množine						
	Úspešnosť	Presnosť nárastu	Pokrytie nárastu	Presnosť poklesu	Pokrytie poklesu	AUC
RF	0.5416	0.71	0.48	0.41	0.65	0.6413
GNN	0.5416	0.59	0.59	0.48	0.48	0.5573
SVC	0.5208	0.62	0.46	0.44	0.60	0.6357
KNN	0.5208	0.57	0.63	0.44	0.38	0.4417
SVCL	0.5833	0.74	0.55	0.44	0.65	-

Ako je možné pozorovať, výsledky sú veľmi blízko náhodnej predpovedi. Jednotlivé kategórie dokázali v niektorých prípadoch dosiahnuť lepšiu úspešnosť ako celá nami zostavená dátová množina. Podobný výsledok deklarujú aj autori [4][6][7], ktorí sa zamerali na rovnakú predpoveď budúcej ceny ako my. Ako sme však už v časti analýzy mohli postrehnúť jednotlivé kategórie nárastu a poklesu boli výrazne prepojené a nedali sa lineárne oddeliť v žiadnom z našich pozorovaní. Z toho dôvodu podrobíme jeden z modelov (RF) analýze, kedy sa pozrieme na chyby daného modelu.



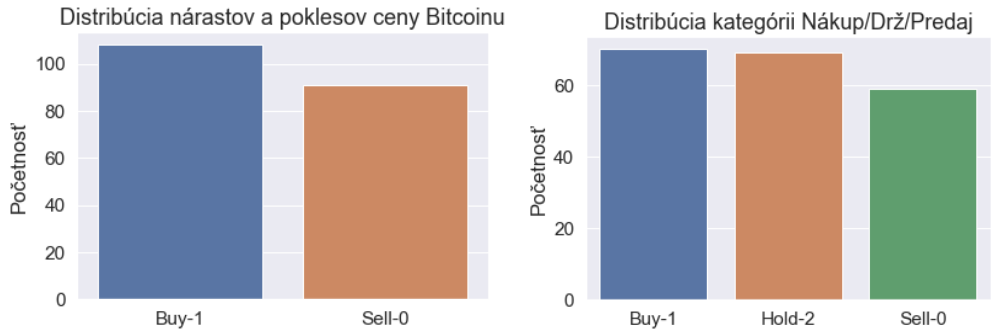
Obrázok 31. Miera a rozloženie chýb modelu náhodný les

Vyššie uvedený výsledok (obrázok 31) pozostáva z chýb piatich po sebe nezávisle natrénovaných modeloch náhodného lesa. Ako je možné vidieť, modely dokážu relatívne

presne určiť smer ceny pri vysokých výkyvoch. Z tohto dôvodu sa potvrdzuje hypotéza, že úspešnosť modelov je skreslená, ich neschopnosťou správne určiť jemné výkyvy cenovej hladiny. Z tohto zistenia môžeme vyvodiť záver, kedy je vhodné pridať ďalšiu triedu pre zmeny ceny v nízkej hladine. V tomto momente prechádzame zo stratégie Nákup/Predaj (Buy/Sell) na stratégiu Nákup/Predaj/Stagnácia (Buy/Sell/Hold), kedy trieda stagnácie bude obsahovať pozorovania s nízkou mierou nárastu alebo poklesu.

4.8.4. Výsledok Experimentu – klasifikácia do 3 tried

Ako sme v predchádzajúcej kapitole preukázali, pridanie tretej kategórie pre jemné výkyvy by mohlo mať za následok, zvýšenie úspešnosti jednotlivých modelov. Tento proces sme vykonali so zohľadnením zachovania distribúcie jednotlivých tried. Z tohto dôvodu bola hranica pre túto kategóriu stanovená v intervale $< -100{,}100 >$ dolárov. Rozloženie kategórií pred a po zavedení tretej kategórie môžeme vidieť na obrázku 32 nižšie.



Obrázok 32. Početnosť jednotlivých kategórií

Tabuľka 12. Metriky modelov pri využití čŕt z celej dátovej množiny (klasifikácia do 3 tried)

Výsledky na celej dátovej množine				
	Úspešnosť	Presnosť/Pokrytie Nákup	Presnosť/Pokrytie Predaj	Presnosť/Pokrytie Stagnácia
RF	0.5000	0.57 / 0.38	0.38 / 0.36	0.52 / 0.85
GNN	0.4166	0.54 / 0.32	0.36 / 0.42	0.38 / 0.57
SVC	0.4583	0.43 / 0.65	0.50 / 0.14	0.50 / 0.50
KNN	0.4375	0.29 / 0.54	0.50 / 0.13	0.60 / 0.60
SVCL	0.4166	0.37 / 0.37	0.64 / 0.45	0.27 / 0.44

Ako môžeme vidieť v tabuľke 12Tabuľka 12 úspešnosť síce klesla ale v kontexte pridania novej triedy sme od náhodného rozdelenia vzdialenejší. Vzhľadom na metriky presnosti a pokrytia, model vykazuje podobné výsledky ako pri dvoch triedach.

4.8.5. Zhodnotenie predbežného experimentu

Nedostatok riešenia v aktuálnom stave spočíva v nedostatku dát. Tento stav nepovažujeme za optimálny, preto by sme smerovanie nášho výskumu sústredili na úroveň hodinových predpovedí. Týmto prístupom naša dátová množina narastie 24 násobne, čím by sa mohol tento problém eliminovať. Taktiež by sme týmto prístupom overili vplyv posunu cieľového atribútu bližšie k sledovaným dátam, nakoľko by sme mohli definovať cieľový atribút ako vývoj ceny v nadchádzajúcej hodine. Ďalšou možnosťou by mohla byť transformácia úlohy na regresnú. Práca môže ponúknuť porovnanie poskytnutých metód a zistiť ich výkonnosť pri rôznych obchodných stratégiách. Kombináciou oboch riešení môže byť taktiež využitie ARIMA modelov. Táto transformácia dát však bude podmienená redukciou sledovaných kategórií, nakoľko vo viacerých črtách sa nedá hodinová granularita pozorovaní dosiahnuť, prípadne ich zmysluplnosť by bola otázna.

4.9. Experiment – predikcia posledného známeho pohybu ceny

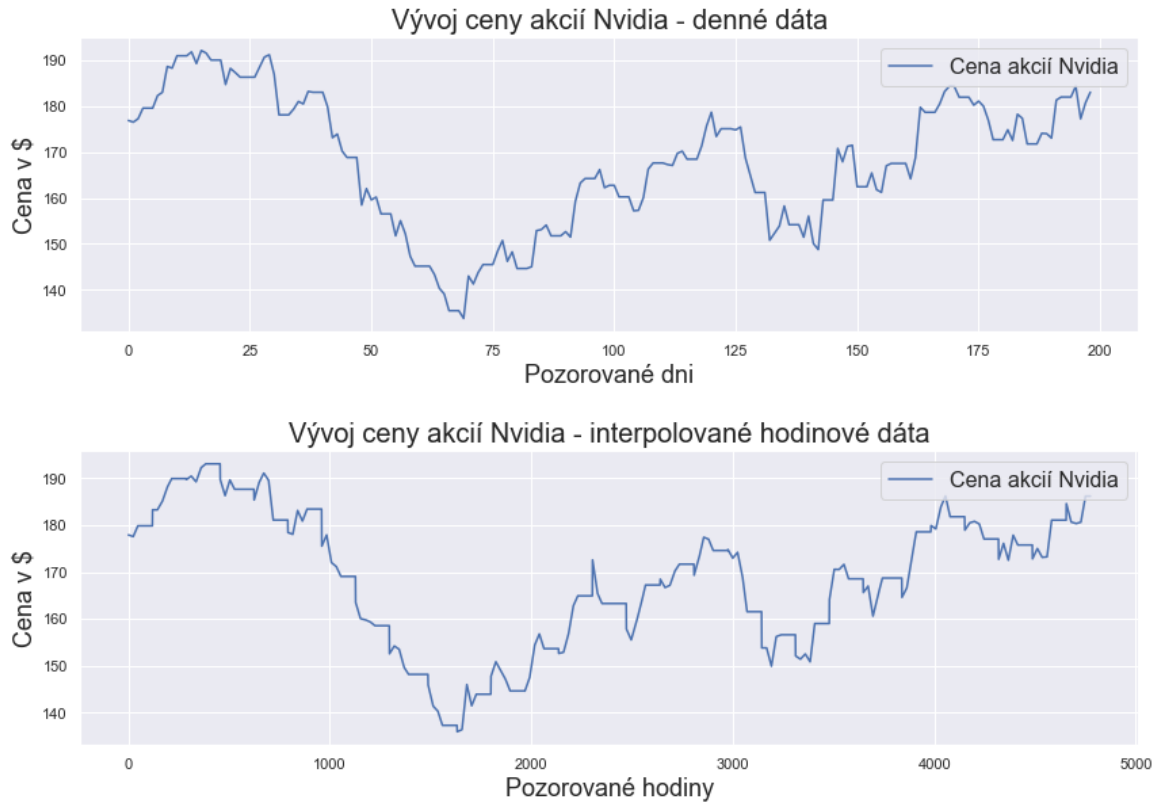
V predbežnom experimente sme preukázali úspešnosť okolo 55% pre triedy v prípade binárnej klasifikácie a 45% pri klasifikácií do 3 tried. Časové rady na trhu kryptomien sa pritom môžu riadiť jednoduchšou logikou. Ak jednotlivé výkyvy trhu majú dlhodobejší trend, adekvátny prístup môže byť aj predikcia budúcej hodnoty na základe predchádzajúceho pohybu. Preto prevedieme jednoduchý test a skonštruujeme model, ktorého úlohou bude predikovať vývoj na základe posledného známeho pohybu ceny Bitcoinu. Tento experiment bude mať aj za úlohu odhaliť či sa modely využité v experimentoch nenaučili len predikovať poslednú známu hodnotu, nakoľko v teoretickej rovine predstavuje hodnotu najbližšie položenú výslednej hodnote. Tento test bude prevedený na dennej aj hodinovej báze.

4.9.1. Predspracovanie a úprava dát

Nakoľko oproti predbežnému experimentu nebudeme pracovať so základnou dátovou množinou je nutné dáta upraviť na požadovanú hodinovú granularitu. To predstavuje vo väčšine prípadov zmeniť agregačnú hodnotu z dňovej na hodinovú. Avšak niektoré zdroje ako napríklad historické dáta vývoja akcií Nvidia či Intel sa nám nepodarilo získať. Z toho dôvodu, ak sme nechceli tieto dáta vylúčiť museli sme pristúpiť k doplneniu chýbajúcich atribútov. Tento proces však predstavoval doplnenie zvyšných 6700 chýbajúcich dát. Zvolili sme preto proces lineárnej interpolácie, kedy chýbajúce hodnoty boli vypočítavané na základe vzťahu nižšie. Kde $v_{d_{tn}}$ predstavuje cieľovú cenu komodity daného dňa *d* v čase *n*.

$$v_{d_{tn}} = \mathrm{v}_{\mathrm{d}_{\mathrm{to}}} + n * \frac{\left(\mathrm{v}_{\mathrm{d+1}_{\mathrm{to}}} - \mathrm{v}_{\mathrm{d}_{\mathrm{to}}}\right)}{24}$$

Týmto prístupom sme dokázali dopočítať chýbajúce dáta bez využitia tradičných prístupov pomocou priemerov či mediánov, ktoré by v tomto prípade znamenalo zanesenie šumu do dát. Na obrázku 33 nižšie môžeme vidieť dáta na dňovej báze (300 pozorovaní) v porovnaní s interpolovanými dátami (7100 pozorovaní).



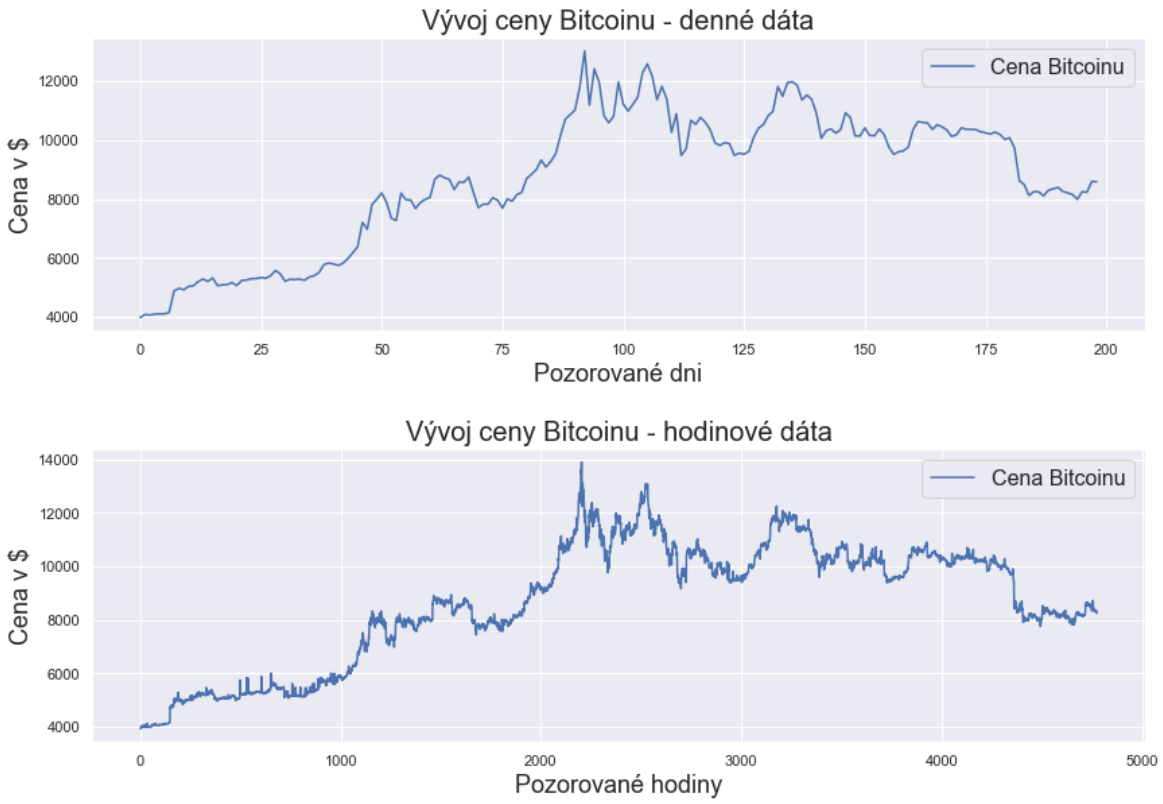
Obrázok 33. Porovnanie denných a interpolovaných hodinových dát

Ako je možné vidieť výsledná cena si uchováva svoj trend. Čo ale môžeme pozorovať sú predĺžené priame línie, ktoré sú spôsobené dňami, ktoré boli chýbajúce už v dátovej množine na dennej báze. Tieto hodnoty vtedy boli nahradené poslednou známou hodnotou, nakoľko išlo prevažne o dni, kedy burza nebola otvorená a interpolácia ceny by bola zavádzajúca. Z toho dôvodu sa tieto úseky predĺžili.

4.9.2. Zmena cieľového atribútu

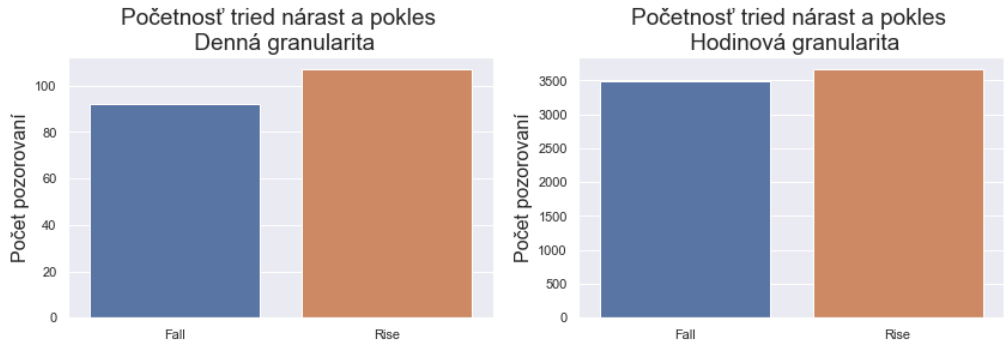
Ako sme uviedli cieľový atribút pre nás predstavuje zatváracia cena nasledujúceho obdobia. V prípade ceny Bitcoinu sme museli zmeniť zdroj, z ktorého sme dáta získavali nakoľko pôvodný neposkytoval hodinovú granularitu. Taktiež usudzujeme, že interpolácia cieľového

atribútu by predstavovala výrazné uľahčenie úlohy pre modely, nakoľko by väčšina dát získala čisto stúpajúci alebo klesajúci trend a nereflektovala by chcený stav trhu. Z toho dôvodu sme sa rozhodli získať nové hodinové dáta, ktoré sme podrobili analýze voči naším pôvodným dátam. Na obrázku 34 je možné vidieť reálne dáta na dennej a hodinovej úrovni.



Obrázok 34. Porovnanie denných a hodinových cien Bitcoinu

Ako je možné vidieť trend ceny Bitcoinu je plne zachovaný ale obsahuje výrazný šum. Týmto prístupom však môže byť porušená aj distribúcia cieľového atribútu nárastu respektíve poklesu. Preto na obrázku 35 zobrazujeme zmenu rozdelenia do týchto tried.



Obrázok 35. Porovnanie početnosti tried nárastu a poklesu v denných a hodinových pozorovaniach

Na základe tohto porovnania môžeme konštatovať, že jednotlivé triedy sa ešte viac týmto prístupom vyvážili čo prispeje k relevantnejším výsledkom v nasledujúcich experimentoch. Taktiež máme za preukázané, že dodatočná exploratívna analýza nie je nutná, nakoľko z vyššie uvedených prípadov sa trendy a rozloženia dát výrazne nemenia.

4.9.3. Výsledky experimentu

Nižšie uvádzame výsledky na jednotlivých dátových množinách dosiahnuté jednoduchou predpoveďou posledného známeho pohybu ceny Bitcoinu.

Tabuľka 133. Výsledky metódy predpovedania posledného známeho pohybu – binárny prístup

Výsledky binárnej klasifikácie						
	Úspešnosť	Presnosť nárastu	Pokrytie nárastu	Presnosť poklesu	Pokrytie poklesu	AUC
LVP	0.4286	0.44	0.44	0.41	0.41	0.4283

Tabuľka 144. Výsledky metódy predpovedania posledného známeho pohybu - klasifikácia do troch tried

Výsledky klasifikácie do troch tried							
	Úspešnosť	Presnosť nárastu	Pokrytie nárastu	Presnosť poklesu	Pokrytie poklesu	Presnosť stagnácie	Pokrytie stagnácie
LVP	0.3174	0.28	0.28	0.26	0.26	0.41	0.41

Ako môžeme vidieť jednotlivé metriky nevykazujú ani náhodnú úspešnosť. Tento fakt nám prezrádza, že neexistuje krátkodobý trend, na základe ktorého by sa dalo odhadnúť smerovanie vývoja ceny Bitcoinu. Vylúčením tohto faktu sme sa ubezpečili, že modely sa neriadia pri svojich predikciách triviálnou logikou.

5.2.Predikcia na hodinovej báze – bežné prístupy

V závere predbežného experimentu sme naznačili smerovanie tejto časti. Hlavným cieľom tohto experimentu bude eliminácia, prípadne redukcia nepriaznivých faktorov. Tieto faktory predstavujú primárne veľkosť dátovej množiny a umiestnenie cieľového atribútu celých 24 hodín, za nami sledované obdobie. Oba tieto problémy sme sa rozhodli vyriešiť prechodom z dennej predikcie zatváracej ceny Bitcoinu na hodinovú predikciu. Týmto experimentom preveríme efektivitu bežných prístupov.

5.2.1. Definícia problému

Samotná definícia problému, zostáva oproti predbežnému experimentu nemenná. Cieľom experimentu je predikcia budúceho vývoja cenovej hladiny Bitcoinu s hodinovou granularitou. V tomto experimente sa však na problém pozrieme z dvoch hľadísk a to ako :

- regresný problém,
- klasifikačný problém.

Okrem klasifikačných modelov sa v tomto experimente zameriame aj na regresné metódy, ktoré budú pre prípad klasifikácie spätne mapované do tried s nulovým prahom prechodu. Tento prístup volíme z dôvodu základného porovnanie klasifikačných a regresných modelov. Pre regresné metódy budú prevedené taktiež osobitné porovnania s vlastnými metrikami. Základom však ostáva overenie výsledkov binárnej klasifikácie do tried nárastu a poklesu so zameraním sa na jednotlivé modely a ich parametre. Taktiež overíme hypotézu troch tried, ktorá sa v prípade dennej predikcie javí ako úspešný proces riešenia problému slabých výsledkov binárnej klasifikácie. Pre všetky základné modely bude prevedené komplexné hľadanie optimálnych parametrov, ktoré bude zaznamenané do dátovej množiny pre serializáciu výsledkov a následnú možnosť filtrácie výsledných modelov na základe výsledkov v jednotlivých metrikách. Týmto postupom bude možné vybrať rôzne prístupy pre rôzne obchodné stratégie, čím naša práca môže slúžiť ako podklad pre optimalizáciu obchodných pravidiel pri tvorbe vhodnej stratégie na danom trhu.

5.2.2. Proces optimalizácie modelov – regresný problém

Pri regresných modeloch sme volili dva základné prístupy. Lineárnu regresiu a ARIMA modely. Pri lineárnej regresií sme pristúpili aj k online trénovaniu, kedy proces trénovania je vždy realizovaný zo všetkých predchádzajúcich historických pozorovaní. Týmto prístupom sa zvyšuje presnosť modelu, nakoľko dokáže počítať vždy s najaktuálnejšími dátami. Výstupom tohto hľadania sú hodnoty metrík MAE, MSE a R^2 spolu s jednotlivými parametrami modelov vo výstupnej dátovej množine.

5.2.3. Vyhodnotenie regresného prístupu

Výsledky najlepších modelov podľa metriky MSE môžeme vidieť v tabuľke 15. Na základe týchto výsledkov vieme modely výkonnostne porovnať medzi sebou. Avšak pri nami definovanom probléme je tento prístup nie úplne ideálny. Preto sme sa rozhodli regresné riešenie spätne mapovať do tried, čím dokážeme regresnými prístupmi riešiť aj klasifikačný problém. Prah prechodu sme ponechali rovnaký ako pri úvodnom triedení cien do tried

nárastu a poklesu aj keď si uvedomujeme, že práve posun tohto prahu prechodu vo výsledku by mohol ešte viac ovplyvniť úspešnosť regresných riešení.

Tabuľka 15. Výsledné metriky regresných modelov

Výsledky binárnej klasifikácie				
Model	Parametre	MSE	MAE	R^2
ARIMA	p: 0, d: 0, q: 1	1.347162	0.280134	0.137253
LR	norm.: true, hist.: 0	1.389043	0.297454	0.110431
LRo	norm.: false, hist.: 0	1.381433	0.293383	0.115304

5.3. Proces optimalizácie modelov – klasifikačný problém

Tak ako aj v prvom experimente sme zvolili prvotný prístup cez tradičné klasifikačné modely pomocou ktorých sme klasifikovali triedy nárastu / poklesu prípadne nárastu / stagnácie / poklesu. Konkrétne išlo o :

- metódu náhodný les,
- k – najbližších susedov,
- metódu podporných vektorov
- naivný Bayesov klasifikátor.

Okrem týchto modelov sme však pristúpili aj k regresným riešeniam (Lineárna regresia, ARIMA modely), ktoré sme na základe výsledku s nulovým prahom prechodu opätovne mapovali do vyššie spomenutých tried, čím sme dosiahli porovnateľnosť týchto modelov. Pre každý model bolo skonštruované mriežkové vyhľadávanie parametrov aby sme dokázali vybrať modely s najadekvátnejším nastavením parametrov. Týmto prístupom sme vybudovali dve dátové množiny výsledkov, ktoré obsahujú výsledky takmer 3000 rôznych modelov. Kompletné zoznamy parametrov pre modely je uvedený v Prílohe G.

5.3.1. Vyhodnotenie binárnej klasifikácie

Ako sme spomenuli v minulej kapitole, výsledkom hľadania vhodných parametrov sú metriky úspešnosti, presnosti, pokrytia a F1 skóre pre každý jeden model s jeho parametrami. Nakoľko výsledky všetkých modelov sú serializované, poskytuje nám to možnosť filtrovania a vyhľadania najlepšieho modelu pre danú metriku. Tieto výsledky metrík sa dajú následne využiť pri konkrétnej stratégii obchodovania. Z toho dôvodu uvádzame výsledky najlepších modelov vo svojej kategórií.

Tabuľka 16. Výsledky bežných modelov – binárna klasifikácia

Výsledky binárnej klasifikácie					
Model	Úspešnosť	Presnosť nárastu	Pokrytie nárastu	Presnosť poklesu	Pokrytie poklesu
ARIMA	0.606615	0.600240	0.688705	0.615646	0.520863
LR	0.589437	0.592497	0.630854	0.585781	0.546110
LRo	0.585503	0.598846	0.571625	0.572802	0.600000
GNB	0.524930	0.525794	0.725436	0.523718	0.313985
KNN	0.503871	0.511721	0.626095	0.494838	0.478023
SVC	0.554117	0.560499	0.626095	0.546083	0.478023
RF	0.582958	0.575506	0.711385	0.597800	0.449172

5.3.2. Vyhodnotenie klasifikácie do troch tried

Týmto prístupom sa snažíme overiť hypotézu na základe predbežného experimentu, kde sa nám profilovala hypotéza, že modely nadobúdajú istotu svojich predikcií práve v najväčších výkyvoch. Nevýhodou výsledku však bola malá testovacia množina a práve týmto testom môžeme overiť výsledok predbežného experimentu. Z tohto dôvodu bude prevedený analogický prístup ako pri binárnej klasifikácií s príslušnou serializáciou výsledkov ako v predošlom prístupe. V tabuľke nižšie môžeme vidieť najlepšie modely s ich výsledkami, ktoré sa nám podarilo vytrénovať.

Tabuľka 17. Výsledky bežných modelov –klasifikácia do troch tried

Výsledky klasifikácie do troch tried							
Model	Úspešnosť	Presnosť/Pokytie nárastu		Presnosť/ Pokrytie poklesu		Presnosť/Pokrytie stagnácie	
ARIMA	0.409571	0.4162	0.1814	0.4656	0.1380	0.4016	0.8330
LR	0.392254	0.4145	0.2146	0.4563	0.1065	0.3813	0.7836
LRo	0.396904	0.4238	0.1415	0.4806	0.1402	0.3838	0.8311
GNB	0.388653	0.3894	0.2584	0.3581	0.2018	0.3970	0.7033
KNN	0.385644	0.3603	0.3226	0.3470	0.3340	0.4380	0.5000
SVC	0.424349	0.4164	0.3884	0.4351	0.2500	0.4269	0.6344
RF	0.445813	0.4322	0.4294	0.4650	0.3074	0.4474	0.5953

Ako môžeme vidieť výsledky zaznamenali jemný prepad, čo môže byť spôsobené väčšou dátovou množinou. Taktiež môžeme vidieť, že najúspešnejším modelom zostáva metóda náhodného lesa a to aj v metrikách presnosti jednotlivých kategórií.

5.4. Vlastný predikčný model

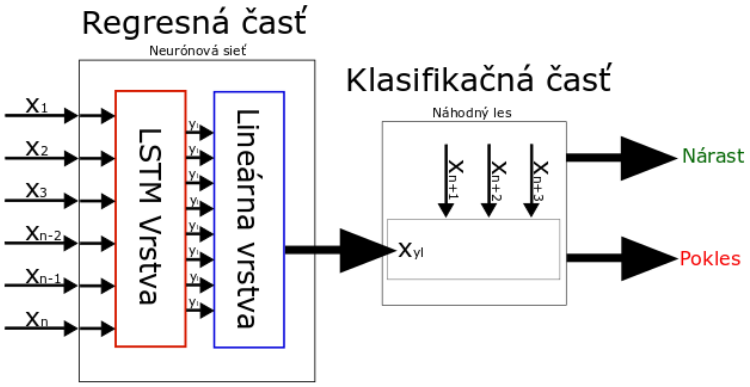
V predchádzajúcich experimentoch sme využili bežne dostupné a relatívne jednoducho implementovateľné metódy. Z uvedeného dôvodu sme sa rozhodli pre implementáciu vlastného modelu, ktorý budeme schopný porovnať s bežnými prístupmi. Model bude pozostávať z dvoch častí. Konkrétne sa bude jednať o regresnú časť pozostávajúcu z rekurentnej neurónovej siete, ktorej úlohou bude odhadnúť vývoj a klasifikačnej ktorá nastaví vhodný prah prechodu do jednotlivých tried - nárastu alebo poklesu. Model bude následne trénovaný rôznymi prístupmi za využitia rôznych prístupov. Vyhodnotíme optimálne parametre trénovania každej časti modelu a prevedieme porovnanie úspešnosti oproti ostatným použitým metódam.

5.5. Konštrukcia modelu – regresná časť

Ako sme spomenuli model bude pozostávať z dvoch častí (obrázok xx). Neurónová sieť bude skonštruovaná z dvoch vrstiev, konkrétne pôjde o LSTM vrstvu, ktorá dokáže zachytávať a sledovať historické zmeny. Nevýhodou LSTM vrstvy však je, že využíva výstupnú aktivačnú funkciu tangens, ktorej hodnoty nadobúdajú interval $(-1.0\ ,\ 1.0)$. Práve preto sme normalizovali dáta na úrovni trénovacej časti dátovej množiny na interval $(-0.5\ ,\ 0.5)$. Týmto prístupom sme zabezpečili, že dáta určené pre trénovanie nenesú v sebe informáciu o budúcom trende vývoja atribútov. Znížený interval normalizácie slúži na variabilitu modelu reagovať aj na hodnoty nad prah normalizácie a pritom teoreticky zostať v intervale výstupných hodnôt daného modelu. Lineárna vrstva následne má za úlohu zjemniť prechod k výslednej hodnote, prípadne reagovať na situáciu kedy by výsledná hodnota z LSTM bola limitovaná práve jej intervalom.

5.6. Konštrukcia modelu – klasifikačná časť

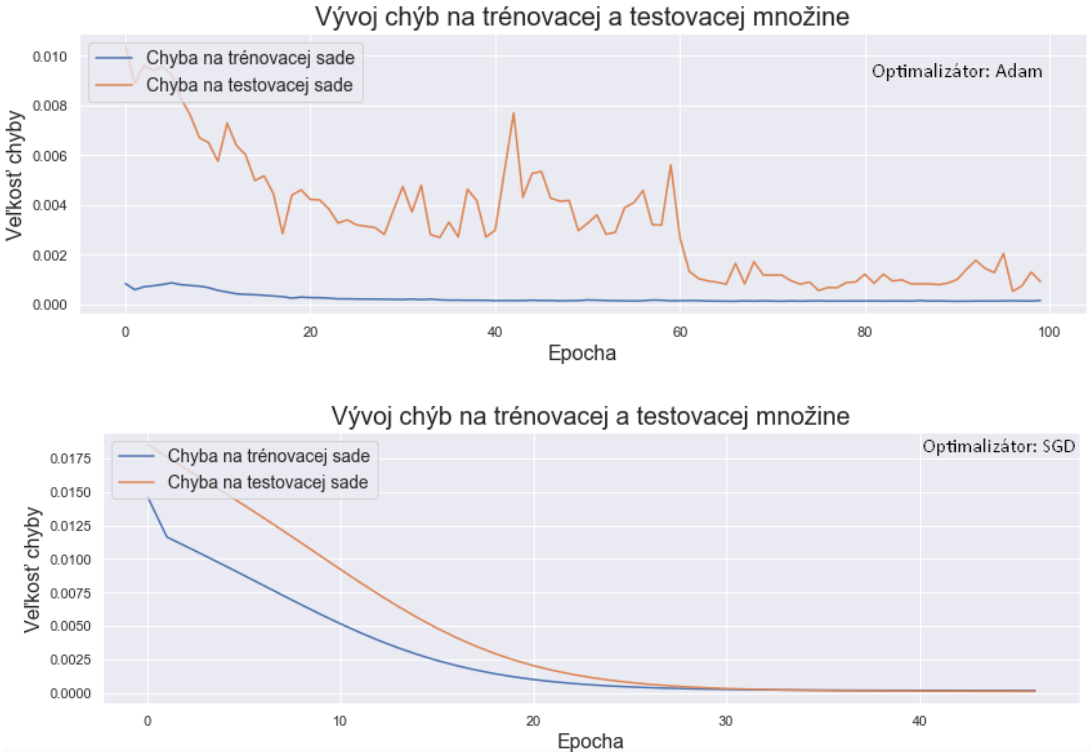
Klasifikačná časť bude realizovaná klasifikátorom náhodný les, kde bude zodpovedný pre určenie prahu prechodu do tried nárastu/poklesu prípadne nárastu/stagnácie/poklesu. Tento prístup nám zároveň umožňuje dodatočne doplniť do sady dát črty, ktoré chceme zvýrazniť pri finálnom rozhodovaní. V našom prípade budú tieto dodatočné črty predstavovať črty sentimentu zo sociálnej siete Twitter. Taktiež bude prevedené kompletné vyhľadávanie parametrov.



Obrázok 36. Architektúra vlastného modelu

5.7.Trénovanie modelu

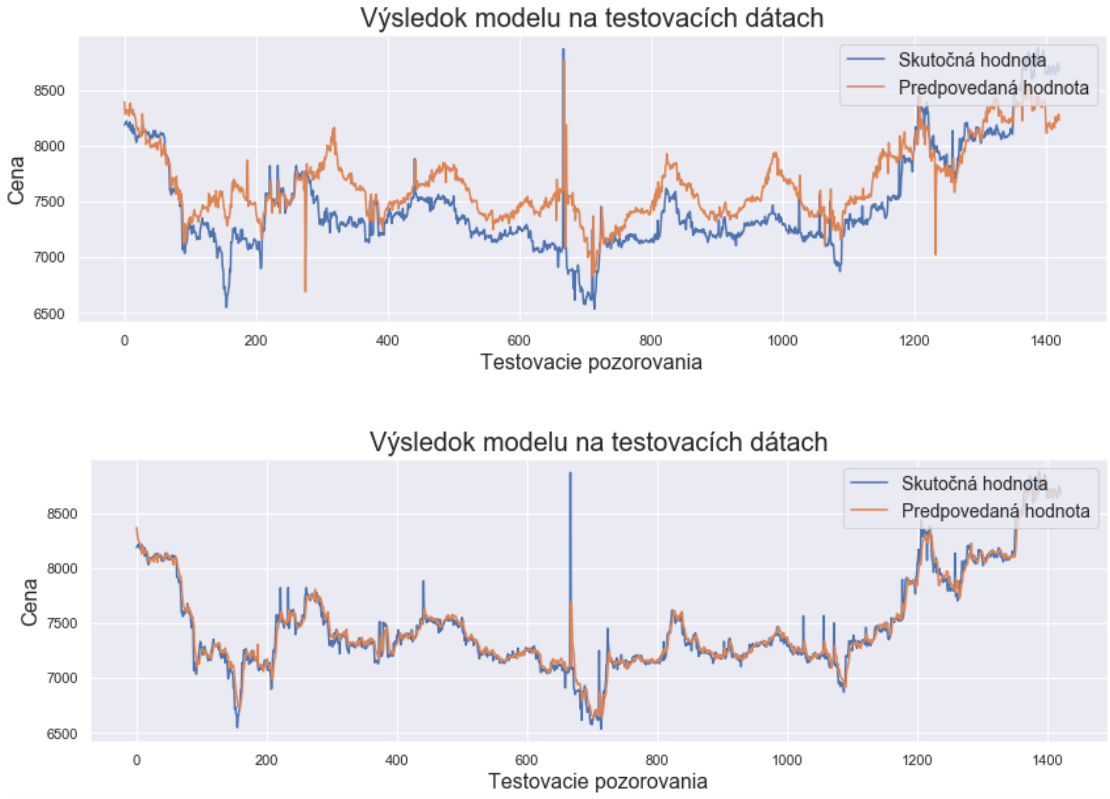
Výsledný model sme podrobili úvodnému trénovaniu na 100 epochách, aby sme zistili jeho reakciu a vývoj chybovej funkcie na trénovacích a testovacích dátach. Odskúšané boli dva optimalizačné prístupy. Konkrétne sa jednalo o ADAM a SGD optimalizátor. Na obrázku nižšie môžeme pozorovať vývoj chýb na trénovacej a testovacej sade. Z obrázku vyplýva, že SGD optimalizátor je rezistentnejší na pretrénovanie a práve z toho dôvodu sa javí ako vhodný prístup pri trénovaní výsledného modelu.



Obrázok 37. Vývoj chýb na trénovacej a testovacej sade pri použití optimalizátorov Adam a SGD

5.8. Výsledok trénovania neurónovej siete

Odpozorovaním z obrázka x vieme určiť efektívny počet epoch, počas ktorých má zmysel model trénovať. Pri každom trénovaní bola taktiež prevedená kontrola trénovacej a testovacej chyby, či náš model nie je náhodou preučený. Rýchlosť učenia bola nastavená na 0.001 pri rozdelení trénovacíh a testovacích dát v pomere 80:20. Jednotlivé výsledky sa získavali takzvaným online prístupom, kedy model predpovedá najviac jednu hodnotu dopredu. Týmto postupom sme zabezpečili predikčnú schopnosť modelu a spravili ho imúnnym voči chybám, ktoré sám produkuje. Na obrázku nižšie môžeme vidieť ako vyzerá predikovaná (testovacia) časť našej dátovej množiny oproti pravdivým dátam, pri využití efektívneho počtu epoch pre ADAM a SGD optimalizátor.



Obrázok 38. Výsledok model na testovacích dátach pri použití optimalizátora Adam a SGD

Z vyššie uvedených zobrazení predikcie je jasné, že SGD optimalizátor sa dokáže tesnejšie natrénovať pri nižšom počte epoch. Avšak prehlásenie lepšieho modelu môžeme previesť až na základe výsledku klasifikačnej časti, ktorú vyskúšame s využitím oboch prístupov optimalizácie a obdobným procesom hľadania parametrov pre klasifikačnú časť náhodného lesa . Nižšie môžeme vidieť tabuľku výsledkov pre 3 najlepšie inštancie nami navrhnutého modelu s jednotlivými hodnotami nasledujúcich parametrov:

- Epochy (e)
- Rýchlosť učenia (lr)
- Optimalizátor (opt)
- Hĺbka náhodného lesa (d)
- Minimálny počet vzoriek pre delenie (mss)
- Minimálny počet vzoriek pre vetvu (msl)

Tabuľka 18. Výsledky vlastného modelu modelov – binárna klasifikácia

Výsledky vlastného modelu - binárna klasifikácia					
Parametre	Úspešnosť	Presnosť nárastu	Pokrytie nárastu	Presnosť poklesu	Pokrytie poklesu
e (30), lr (0.001) , opt(Adam) , d(8) , mss (2) , msl (1)	0.6191	0.6543	0.6031	0.5852	0.6373
e (30), lr (0.001) , opt(Adam) , d(8) , mss (2) , msl (1)	0.6180	0.6536	0.6010	0.5841	0.6373
e (30), lr (0.001) , opt(Adam) , d(8) , mss (2) , msl (1)	0.6171	0.6526	0.6010	0.5832	0.6355

Ako je možné vidieť model dosahuje mierne lepšie výsledky oproti tradičným metódam. Prekvapujúcim môže byť však počet epoch ako aj samotný optimalizátor najlepších modelov. Napriek tomu, že model dosahuje lepšie výsledky je nutné sa zamyslieť, či dané zlepšenie vzhľadom na komplexnosť modelu má praktické využitie. Pre porovnanie nižšie uvádzame najlepšie modely pre klasifikáciu do troch tried.

Tabuľka 18. Výsledky vlastného modelu modelov –klasifikácia do troch tried

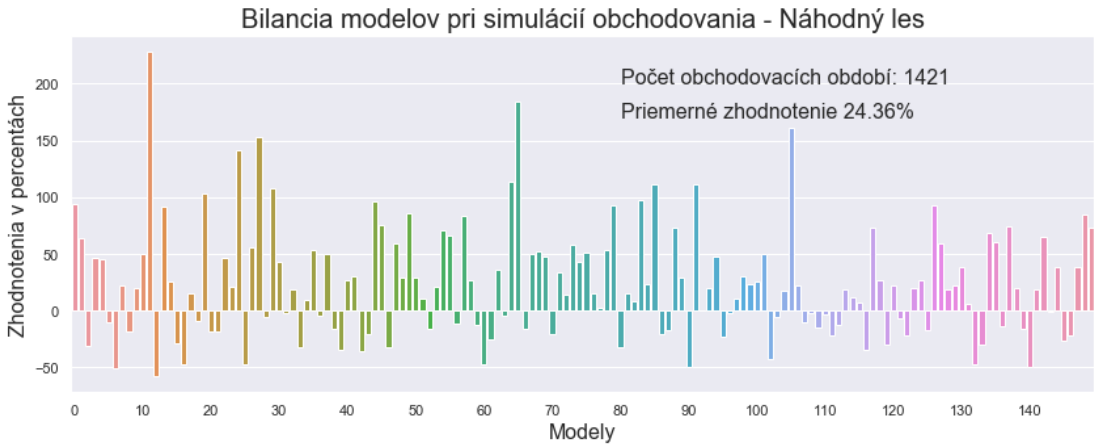
Výsledky vlastného modelu – klasifikácia do troch tried				
Parametre	Úspešnosť	Presnosť /Pokrytie nárastu	Presnosť /Pokrytie poklesu	Presnosť /Pokrytie stagnácie
e (30), lr (0.001) , opt(Adam) , d(8) , mss (2) , msl (1)	0.4084	0.3992/ 0.3526	0.4044/ 0.3214	0.4162/ 0.5323
e (30), lr (0.001) , opt(Adam) , d(8) , mss (2) , msl (1)	0.4070	0.3985/ 0.3491	0.3958/ 0.3053	0.4181/ 0.5453
e (30), lr (0.001) , opt(Adam) , d(8) , mss (2) , msl (1)	0.4061	0.3943/ 0.3385	0.4007/ 0.3285	0.4165/ 0.5323

5.9. Vyhodnotenie modelov na úrovni simulácie

Metriky uvedené vyššie nám síce umožňujú ohodnotiť model na úrovni úspešnosti, presnosti či pokrytia, avšak v reálnom prostredí môže byť stále model nepresný, nakoľko nerozlišuje váhu svojich rozhodnutí. Rozhodli sme sa skonštruovať jednoduchý testovací scenár, ktorý na základe jednoduchých obchodovacích pravidiel bude spravovať dané obchodné portfólio. Test sa bude riadiť výstupmi z modelu a bude zohľadňovať dopad jeho rozhodnutí na reálnych dátach. Aj keď je tento prístup veľmi jednoduchý môže nám pomôcť odhaliť či má náš model predpoklad byť úspešný pri obchodovaní na danom trhu. Pre testovanie v tomto scenári sme vybrali metódu náhodného lesa, nakoľko je to pravdepodobnostný model a vykazoval jedny z najlepších výsledkov. Ďalším bude nami navrhnutý model. Pravidlá simulátora pre obchodovanie boli zadefinované nasledovne:

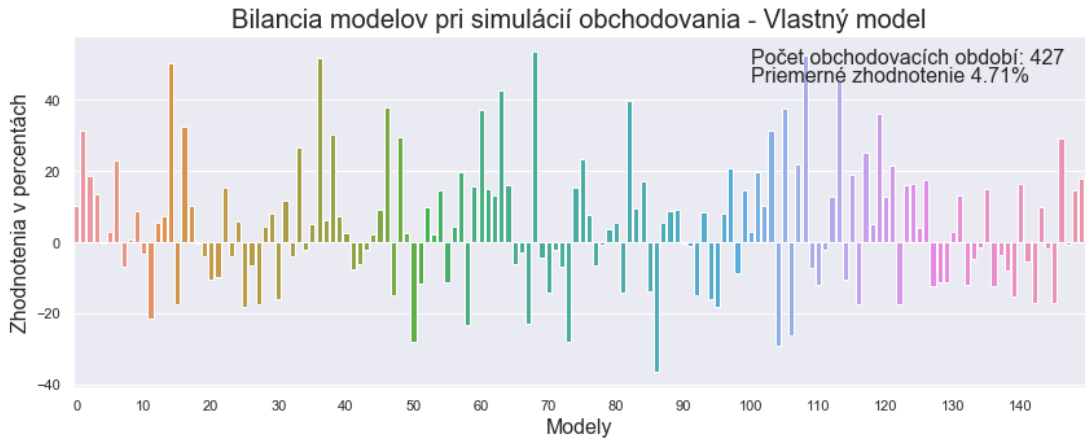
- simulátor disponuje portfóliom 10000€ v mene BTC,
- simulátor disponuje hotovosťou 5000€,
- každú hodinu model prevádza iba jednu operáciu na základe výsledku modelu,
- simulátor v jednej operácií môže narábať iba s hodnotou 500€,
- ak simulátor nedisponuje hotovosťou drží portfólio po najbližšiu predikciu poklesu,
- ak simulátor nedisponuje prostriedkami v portfóliu čaká na predikciu nárastu.

Aby bol výsledok relevantnejší vytrénovali sme nezávisle 150 modelov (s rôznymi sekvenciami trénovacích a testovacích dát). Obrázok nižšie znázorňuje bilanciu jednotlivých modelov. Ako môžeme vidieť väčšina modelov dosahuje kladné hodnoty čo znamená, že sa im aj napriek relatívne nízkej úspešnosti (60%) darí predpovedať výkyvy a pri jednoduchej obchodnej stratégií postupne zhodnocovať zverený kapitál. Priemerne modely dosahujú 24,36% zhodnotenie úvodného vkladu.



Obrázok 38. Efektivita modelu náhodného lesa so zvereným kapitálom

Nižšie na obrázku uvádzame výsledok toho istého testu pre vlastný model. Nevýhodou však je, že nakoľko regresná časť modelu vytvára vstupy pre klasifikačný model dátová množina, ktorú máme k dispozícii, nezabezpečuje adekvátne množstvo dát na ktorom by sa dal model plnohodnotne overiť. Ako môžeme sledovať výsledky nášho modelu sú horšie (4.71%) v porovnaní s najlepším modelom náhodného lesa. Príčinou môže byť nedostatok dát na dotrénovanie klasifikačnej časti modelu prípadne nevhodná obchodná stratégia pre tento model. Určite sa za výsledkom ovplyvnilo aj kratšie časové obdobie (počet pozorovaní) v ktorom daný model prevádzal svoje operácie nákupu a predaja.



Obrázok 35. Efektivita vlastného modelu pri práci so zvereným kapitálom

Z tohto dôvodu nemôžeme s určitosťou prehlásiť, ktorý z modelov je prakticky výkonnejší, avšak výsledok môže poskytovať základ pre budúcu prácu s týmito modelmi a procesom optimalizácie obchodných pravidiel na základe výsledkov daných modelov. Taktiež možným smerovaním ďalšieho pozorovania môže byť zohľadnenie istoty modelov pri svojich predikciách, pričom sa môže taktiež upraviť stratégia obchodovania priamo pre konkrétny model. V oboch prípadoch však modely majú vyššiu tendenciu zhodnocovať zverený kapitál, čím sa definujú ako prakticky použiteľné.

6. Zhodnotenie a záver

V tejto práci sme sa venovali predikcií cenovej hladiny Bitcoinu. Práca poskytuje kompletný prehľad o segmentoch ovplyvňujúce tento trh, pričom sme sa zamerali na komplexnú analýzu týchto činiteľov a vykonali predbežný experiment, ktorý odhalil zmysluplnosť jednotlivých kategórií. Úvod práce je venovaný budovaniu teoretického slovníka a pojmov s ktorými by mal byť čitateľ stotožnený. Rozsiahlou časťou práce je analýza jednotlivých sledovaných segmentov, pričom bol kladený dôraz na schopnosť danej kategórie ovplyvniť budúci vývoj ceny Bitcoinu. Analýzou sa vyprofilovali ako najsignifikantnejšie črty práve z oblasti kryptomien a sentimentu zo sociálnej siete Twitter. Predbežným experimentom na dennej báze sme dosiahli výsledky porovnateľné s autormi [4,6,7], čo predstavuje približne 55%. Výsledok predbežného experimentu nám priblížil možnosti zlepšenia jednotlivých predikcií pridaním triedy stagnácie, ktorá sa javila ako prostriedok zlepšenia jednotlivých predikcií, oddelením najmenej dôležitých pohybov cien. Nakoľko však nami pozorované obdobie v predbežnom experimente obsahovalo len 300 denných pozorovaní, rozhodli sme sa prejsť na hodinovú predikciou, čím by sme mohli potvrdiť dosiahnuté výsledky na dennej báze. Týmto prístupom sme overili jednotlivé postupy predbežného experimentu a potvrdili dokonca v prípade binárnej klasifikácie vylepšili dosahované výsledky na úroveň úspešnosti 59%, čo vzhľadom na veľkosť dátovej množiny na báze hodinových pozorovaní zvyšuje relevanciu dosiahnutých výsledkov. Hypotéza troch tried sa však pri hodinovej predikcií neosvedčila. Zväčšením dátovej množiny sa nám tiež vyprofilovala možnosť využitia neurónových sietí. Tento prístup sme implementovali vo vlastnom modeli, ktorý zahŕňa neurónovú sieť zostavenú z dvoch vrstiev (lstm vrstvy a lineárnej vrstvy) a klasifikačnej časti, ktorá je zodpovedná za mapovanie výsledkov neurónovej siete do príslušných kategórií. Náš model dosahuje 61% úspešnosť pri binárnej klasifikácií, čím prekonal najlepšie výsledky bežných modelov. Nakoľko metrika úspešnosti predikcií nereflektuje reálnu použiteľnosť modelu, rozhodli sme sa skonštruovať jednoduchý simulátor obchodovania, na ktorom model náhodného lesa dosiahol pri 1400 testovacích pozorovaniach, zhodnotenie vkladu v priemere o 26%. Nami navrhnutý model dosahuje mieru zhodnotenia %. Tento výsledok však nemusí byť konečný nakoľko nami definované obchodné pravidlá môžu byť pre niektoré modely limitujúce. Táto časť práce bola prevedená na úrovni základného porovnania úspešnosti modelov pri zhodnocovaní vkladu. Zároveň týmto prístupom práca poskytuje široký základ pre budúce skúmanie tejto domény na úrovni optimalizácie obchodných pravidiel pre jednotlivé modely.

Na začiatku tejto práce sme tiež zadefinovali pojem teória efektívneho trhu, ktorá definuje trh a jeho vývoj ako čisto náhodnú veličinu, avšak ako sme v práci preukázali, existujú

činitele, na základe ktorých je možné odhadnúť smerovanie vývoja. Faktom však ostáva, že stále sledujeme obmedzené spektrum týchto činiteľov, pričom aj kryptomeny ako celok patria do globálnej trhovej ekonomiky. V tom prípade je takmer nemožné dosahovať výrazne uspokojivé výsledky. Obdobie, ktoré sme sledovali taktiež nepokrýva celé historické spektrum vývoja Bitcoinu. Teória efektívneho trhu sa v našom prípade javí ako porušená avšak treba poznamenať, že v nami sledovanom časovom spektre málokedy zaznamenal Bitcoin viac ako 3% výkyv pričom v historických obdobiach, ktoré naša dátová množina nepokrýva sú tieto výkyvy v desiatkach percent. Z toho dôvodu môžeme porušenie teórie efektívneho trhu prehlásiť len na nami sledovanom časovom období nie však nad Bitcoinom globálne.

Zdroje

[1] DAVIES G., A history of money: from ancient times to the present day, 3rd ed., with Revisions. Cardiff: University of Wales Press, 2002, ISBN 978-0708317174

[2] NAKAMOTO S., Bitcoin: A Peer-to-Peer Electronic Cash System, s. 9., 2009, Dostupné na: https://bitcoin.org/bitcoin.pdf

[3] FAMA E. F., Portfolio Analysis in a Stable Paretian Market, Management Science., vol. 11, INFORMS, 1965, Dostupné na: www.jstor.org/stable/2628055

[4] LAMON C., NIELSEN E., REDONDO E., Cryptocurrency Price Prediction Using News and Social Media Sentiment, SMU Data Sci. Rev, 2017, Dostupné na: http://cs229.stanford.edu/proj2017/final-reports/5237280.pdf

[5] PRUITT S. W., WHITE R. E., The CRISMA trading system: Who says technical analysis can't beat the market?, The Journal of Portfolio Management, vol. 14, č. 3, s. 55–58, apr. 1988, DOI: https://doi.org/10.3905/jpm.1988.409149

[6] BELL T., Bitcoin Trading Agents, School of Electronics and Computer Science University of Southampton., 2015, Dostupné na: https://www.academia.edu/18279672/Bitcoin_Trading_Agents

[7] MCNALLY S., ROCHE J., CATON S., Predicting the Price of Bitcoin Using Machine Learning, 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), Cambridge, 2018, s. 339–343., ISBN: 978-1-5386-4975-6, DOI: 10.1109/ PDP2018.2018.00060

[8] SAAD M., MOHAISEN A., Towards characterizing blockchain-based cryptocurrencies for highly-accurate predictions, IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Honolulu, HI, 2018, s. 704–709., Electronic ISBN: 978-1-5386-5979-3, DOI: 10.1109/INFCOMW.2018.8406859

[9] ABRAHAM J., HIGDON D., NELSON J., IBARRA J., Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis, SMU Data Science Review, vol. 1, č. 3, s. 22, 2018., Dostupné na: https://scholar.smu.edu/datasciencereview/vol1/iss3/1

[10] QADAN M., NAMA H., Investor sentiment and the price of oil, Energy Econ., vol. 69, s. 42–58, jan. 2018., DOI: 10.1016/j.eneco.2017.10.035

[11] COLIANNI S., ROSALES S., SIGNOROTTI M., Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis, SMU Data Science Review, 2018 s. 5., Dostupné na: https://scholar.smu.edu/datasciencereview/vol1/iss3/1

[12] GUO T., ANTULOV F. N., Predicting short-term Bitcoin price fluctuations from buy and sell orders, IEEE International Conference on Data Mining (ICDM). IEEE, 2018: 989-994, DOI: 10.1109/ICDM.2018.00123

[13] MAO H., COUNTS S., BOLLEN J., Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data, Statistical Finance (q-fin.ST); Computational Engineering, Finance, and Science (cs.CE); Physics and Society (physics.soc-ph), Dostupné na: https://arxiv.org/abs/1112.1051

[14] KAMINSKI J., Nowcasting the Bitcoin Market with Twitter Signals, *Social and* Information Networks (cs.SI), arXiv:1406.7577, 2014, Dostupné na: https://arxiv.org/abs/1406.7577

[15] KIM Y. B., LEE J., PARK N., CHOO J., KIM J.-H., When Bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation, PLOS ONE, vol. 12, 2017, DOI: 10.1371/journal.pone.0177630

[16] RAO T., SRIVASTAVA S., Using Twitter Sentiments and Search Volumes Index To Predict Oil, Gold, Forex and Markets Indices, Proceedings of the 5th Annual ACM Web Science Conference, 2013, ISBN: 978-1-4503-1889-1, DOI: 10.1145/2464464.2464521

[17] MADAN I., SALUJA S., ZHAO A., Automated Bitcoin Trading via Machine Learning Algorithms", Department of Computer Science Stanford University, 2014, Dostupné na: https://www.semanticscholar.org/paper/Automated-Bitcoin-Trading-via-Machine-Learning-Madan/e0653631b4a476abf5276a264f6bbff40b132061

[18] KVASNIČKA, V., BEŇOUŠKOVÁ Ľ, et al., Úvod do teórie neurónových sietí, Iris, 1997, ISBN 8088778301

[19] ROSIC, A.: What is Blockchain Technology? A Step-by-Step Guide For Beginners., [cit. 25.05.2019]. Dostupné na: http://https://blockgeeks.com/guides/what-is-blockchain-technology/

[20] CHOVANCOVÁ a kol.2002, Finančný trh- Nástroje ,transakcie, inštitúcii. Bratislava: Euronion, 2002 14.100-103,s. ISBN 80-88984-31-9.

[21] CIBULA, M.: Definícia strojového učenia. https://smnd.sk/mcibula/, [cit. 25.05.2019]. Dostupné na: https://smnd.sk/mcibula/zakl_info/definicia.html

[22] CIBULA, M.: Lineárna regresia - Algoritmy - Učenie s učiteľom. [cit. 26.05.2019]. Dostupné na: https://smnd.sk/mcibula/alg/linreg.html

[23] CIBULA, M.: Naive Bayes - Algoritmy - Učenie s učiteľom., [cit. 26.05.2019]. Dostupné na: https://smnd.sk/mcibula/alg/NB.html

[24] CIBULA, M.:, M.: Rozhodovacie stromy - Algoritmy - Učenie s učiteľom., [cit. 26.05.2019]. Dostupné na webovej stránke: https://smnd.sk/mcibula/alg/DT.html

[25] KELÍŠEK, A.: Analýza časových radov pomocou neurónových sietí, Zborník príspevkov z konferencie mladých vedeckých pracovníkov Veda a krízové situácie, Ostrava, s. 68 – 73, 2007.

[26] VALÁŠEK, L.: Časové rady a ich dekompozícia, Katedra matematiky a deskriptívnej geometrie STU, 2014, [cit. 09.11.2019], Dostupné na: https://www.math.sk/mpm/wp-content/uploads/2017/11/acr.pdf

[27] MARČEK, D., MARČEK, M., PANČÍKOVÁ, L., Ekonometria a soft computing., Žilina: EDIS, 2008, ISBN 978-80-8070-746-0

[28] MARČEK, D., Ekonometria: Základy, postupy, Aplikačné príklady. Žilina: EDIS, 1999, ISBN 80-7100-557-6

[29] Historické data Bitcoinu, 2019, Dostupné na: https://www.blockchain.com/stats

[30] Dokumentácia k nástroju NLTK, 2019, Dostupné na: https://www.nltk.org

[31] Dokumentácia k metóde StandardScaler, 2019, Dostupné na: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[32] Online burza kryptomien, Dostupné na: https://coinmarketcap.com

Príloha A: Plán práce pre DP I

1. Týždeň	Zhromažďovanie dokumentov.
2. Týždeň	Zhromažďovanie dokumentov.
3. Týždeň	Štúdium literatúry.
4. Týždeň	Štúdium literatúry.
5. Týždeň	Selekcia vhodnej doménovej literatúry.
6. Týždeň	Konzultácia vybranej literatúry.
7. Týždeň	Analýza techník a prístupy riešení.
8. Týždeň	Osnova a štruktúra obsahu DP1.
9. Týždeň	Úvod DP1.
10. Týždeň	Predbežná verzia DP1.
11. Týždeň	Konzultácia a pripomienkovanie DP1.
12. Týždeň	Aplikácia pripomienok a finálna úprava.

Príloha B: Plán práce pre DP II

1. Týždeň	Dokončenie plánu, kompletizácia dát, návrh finálnych čŕt.
2. Týždeň	
3. Týždeň	Finalizácia čŕt a exploratívna analýza.
4. Týždeň	
5. Týždeň	Analýza a prehľadov algoritmov.
6. Týždeň	
7. Týždeň	Analýza techník a prístupy riešení.
8. Týždeň	Návrhy vylepšení
9. Týždeň	Návrh a prevedenie experimentu - prvé výsledky
10. Týždeň	
11. Týždeň	Písanie dokumentu
12. Týždeň	Finalizácia dokumentu a pripomienkovanie

Príloha C: Zoznam čŕt z kategórie Twitter

- tweet_volume – počet tweetov za daný deň
- nonzero_neutral_tweet – počet nenulových neutrálnych tweetov za daný deň
- nonzero_positive_tweet – počet nenulových pozitívnych tweetov za daný deň
- nonzero_negative_tweet – počet nenulových negatívnych tweetov za daný deň
- nonzero_compound_tweet – počet nenulových sumárnych tweetov za daný deň
- sum_positive_mul_follow – Σ (pozitívna časť * počet sledovateľov)
- sum_neutral_mul_follow – Σ (neutrálna časť * počet sledovateľov)
- sum_negative_mul_follow – Σ (negatívna časť * počet sledovateľov)
- sum_compound_mul_follow – Σ (sumárny sentiment * počet followerov)
- sum_positive – Σ (pozitívny sentiment)
- sum_neutral– Σ (neutrálny sentiment)
- sum_negative – Σ (negatívny sentiment)
- sum_compound – Σ (sumárny sentiment)
- avg_sum_positive – Σ (pozitívny sentiment)/ tweet_volume
- avg_sum_neutral – Σ (neutrálny sentiment)/ tweet_volume
- avg_sum_negative – Σ (negatívny sentiment)/ tweet_volume
- avg_sum_compound – Σ (sumárny sentiment)/ tweet_volume
- avg_sum_nonzero_positive - Σ (pozitívna časť)/ nonzero_positive_tweet
- avg_sum_nonzero_neutral - Σ (neutrálna časť)/ nonzero_neutral_tweet
- avg_sum_nonzero_negative - Σ (negatívna časť)/ nonzero_negative_tweet
- avg_sum_nonzero_compound - Σ (sumárny sentiment)/ nonzero_compound_tweet
- avg_sum_positive_mul_follow - avg_sum_positive * počet sledovateľov
- avg_sum_neutral_mul_follow - avg_sum_neutral * počet sledovateľov
- avg_sum_negative_mul_follow - avg_sum_negative * počet sledovateľov
- avg_sum_compound_mul_follow - avg_sum_compound * počet sledovateľov
- avg_sum_nonzero_positive_mul_follow - avg_sum_nonzero_positive * sledovatelia
- avg_sum_nonzero_neutral_mul_follow - avg_sum_nonzero_neutral * sledovatelia
- avg_sum_nonzero_negative_mul_follow-avg_sum_nonzero_negative * sledovatelia
- avg_sum_nonzero_comp_mul_follow - avg_sum_nonzero_compound * sledovatelia

Príloha D: Zoznam čŕt z kategórie webové publikácie

- sum_clheaders_negative – Σ (negatívny sentiment predspracovaných titulkov)
- sum_clheaders_positive– Σ (pozitívny sentiment predspracovaných titulkov)
- sum_clheaders_neutral– Σ (neutrálny sentiment predspracovaných titulkov)
- sum_clheaders_compound– Σ (sumárny sentiment predspracovaných titulkov)
- avg_clheaders_negative– Σ (negatívny sentiment predsp. titulkov)/počet článkov
- avg_clheaders_positive– Σ (pozitívny sentiment predsp. titulkov) /počet článkov
- avg_clheaders_neutral– Σ (neutrálny sentiment predsp. titulkov) /počet článkov
- avg_clheaders_compound– Σ (sumárny sentiment predsp. titulkov) /počet článkov
- sum_headers_negative– Σ (negatívny sentiment titulkov)
- sum_headers_positive– Σ (pozitívny sentiment titulkov)
- sum_headers_neutral– Σ (neutrálny sentiment titulkov)
- sum_headers_compound– Σ (sumárny sentiment titulkov)
- avg_headers_negative– Σ (pozitívny sentiment titulkov) /počet článkov
- avg_headers_positive– Σ (pozitívny sentiment titulkov) /počet článkov
- avg_headers_neutral– Σ (neutrálny sentiment titulkov) /počet článkov
- avg_headers_compound– Σ (sumárny sentiment titulkov) /počet článkov
- sum_clcontents_negative– Σ (negatívny sentiment predsp. obsahu)
- sum_clcontents_positive– Σ (pozitívny sentiment predsp. obsahu)
- sum_clcontents_neutral– Σ (neutrálny sentiment predsp. obsahu)
- sum_clcontents_compound– Σ (sumárny sentiment predsp. obsahu)
- avg_clcontents_negative– Σ (negatívny sentiment predsp. obsahu) /počet článkov
- avg_clcontents_positive– Σ (pozitívny sentiment predsp. obsahu) /počet článkov
- avg_clcontents_neutral– Σ (neutrálny sentiment predsp. obsahu) /počet článkov
- avg_clcontents_compound– Σ (sumárny sentiment predsp. obsahu) /počet článkov
- sum_contents_negative– Σ (negatívny sentiment obsahu)
- sum_contents_positive– Σ (pozitívny sentiment obsahu)
- sum_contents_neutral– Σ (neutrálny sentiment obsahu)
- sum_contents_compound– Σ (sumárny sentiment obsahu)
- avg_contents_negative– Σ (negatívny sentiment obsahu) /počet článkov
- avg_contents_positive– Σ (pozitívny sentiment obsahu) /počet článkov
- avg_contents_neutral– Σ (neutrálny sentiment obsahu) /počet článkov
- avg_contents_compound– Σ (sumárny sentiment obsahu) /počet článkov

Príloha E: Zoznam čŕt z kategórie trh kryptomien

- btccash_close_price
- btccash_market_cap
- d_btccash_open_price
- d_btccash_close_price
- d_btccash_volume
- d_btccash_market_cap
- ethereum_open_price
- ethereum_close_price
- ethereum_volume
- ethereum_market_cap
- d_ethereum_open_price
- d_ethereum_close_price
- d_ethereum_volume
- d_ethereum_market_cap
- litecoin_open_price
- litecoin_close_price
- litecoin_volume
- litecoin_market_cap
- d_litecoin_open_price
- d_litecoin_close_price
- d_litecoin_volume
- d_litecoin_market_cap
- maker_open_price
- maker_close_price
- btccash_volume
- maker_volume
- maker_market_cap
- d_maker_open_price
- d_maker_close_price
- d_maker_volume
- d_maker_market_cap
- monero_open_price
- monero_close_price
- monero_volume
- monero_market_cap
- d_monero_open_price
- d_monero_close_price
- d_monero_volume
- d_monero_market_cap
- xrp_open_price
- xrp_close_price
- xrp_volume
- xrp_market_cap
- d_xrp_open_price
- d_xrp_close_price
- d_xrp_volume
- d_xrp_market_cap

Príloha F: Zoznam čŕt z kategórie trh akcií

- cobalt_price – cena kobaltu
- d_cobalt_price – diferencia cien kobaltu
- oil_price – cena ropy
- d_oil_price – diferencie ceny ropy
- amd_price – cena akcií spoločnosti AMD
- amd_volume – objem obchodovateľných akcií AMD
- d_amd_price – diferencia ceny akcií AMD
- d_amd_volume – diferencia objemu obchodovateľných AMD
- intel_price – cena akcií spoločnosti Intel
- intel_volume – cena obchodovateľných akcií Intel
- d_intel_price – diferencia cien Intel
- d_intel_volume – diferencia obchodovateľných akcií Intel
- nvidia_price – cena akcií spoločností Nvidia
- nvidia_volume – cena obchodovateľných akcií Nvidia
- d_nvidia_price - diferencia ceny akcií Nvidia
- d_nvidia_volume – diferencia obchodovateľných akcií Nvidia

Príloha G: Štruktúra elektronického média

- Data_mining_tools
  - Twitter_API
    - Example of Json
  - News_crawlers
  - Sentiment_extractors
- Data_analyse
  - Twitter_data
  - News_data
  - Historical_data
  - Crypto_data
  - Stockmarket_data
- Experiment
- Data
  - Data.csv
- Documentation