VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INFORMAČNÍCH SYSTÉMŮ
DEPARTMENT OF INFORMATION SYSTEMS

ANALÝZA ČASOVÝCH ŘAD

TIME SERIES ANALYSIS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **SAMUEL BUDAI**
AUTHOR

VEDOUCÍ PRÁCE **Ing. IVANA BURGETOVÁ, Ph.D.**
SUPERVISOR

BRNO 2023

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Zadání bakalářské práce

144247

Ústav:	Ústav informačních systémů (UIFS)
Student:	**Budai Samuel**
Program:	Informační technologie
Specializace:	Informační technologie
Název:	**Analýza časových řad**
Kategorie:	Data mining
Akademický rok:	2022/23

Zadání:

1. Seznamte se s problematikou časových řad a s metodami používanými pro jejich modelování.
2. Po dohodě s vedoucí vyberte metody pro predikci následujících hodnot časových řad, kterými se budete dále zabývat.
3. Seznamte se s dostupnými datovými sadami obsahujícími komunikaci v průmyslových sítích.
4. Navrhněte a implementujte jednoduchou aplikaci, která umožní otestovat vybrané metody na dostupných datových sadách.
5. Vybrané metody otestujte.
6. Zhodnoťte dosažené výsledky.

Literatura:

- JOHNSON, Timothy D. Time Series Analysis with Applications in R, 2nd edition by CRYER, J. D. and CHAN, K.-S. *Biometrics* [online]. Malden, USA: Blackwell Publishing, 2009, ISSN 0006-341X.
- BROCKWELL, Peter J., DAVIS Richard A.: Introduction to Time Series and Forecasting, Springer Cham, 2016, ISBN 978-3-319-29854-2.

Při obhajobě semestrální části projektu je požadováno:

- Body 1 až 3.

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce:	**Burgetová Ivana, Ing., Ph.D.**
Vedoucí ústavu:	Kolář Dušan, doc. Dr. Ing.
Datum zadání:	1.11.2022
Termín pro odevzdání:	17.5.2023
Datum schválení:	20.10.2022

Abstrakt

Táto práca rieši problematiku analýzy časových radov a jej využitie pri detekcii anomálií v priemyselných sieťach. Pre tvorbu modelov predikcie boli v riešení použité algoritmy AR-X, ARIMA, SARIMA, Random Forest, Facebook Prophet a XGB Boost. Okrem toho práca zahŕňa implementáciu algoritmu pre detekciu anomálií z modelov predikcií ako aj riešenie problematiky vysokej sezónnej periódy v prípade algoritmu SARIMA. Vykonaným výskumom sa zistilo, že za použitia vybraných algoritmov je možné predikovať priemyselnú premávku za účelom detekcie, v rámci ktorej sa podarilo odhaliť až 90% útokov. Rovnako práca prináša riešenie vysokej sezónnej periódy za použitia parciálnych časových radov. Tieto výsledky umožňujú experimentálnu integráciu detekcie na základe predikcií do reálnych priemyselných sietí.

Abstract

This thesis deals with the issue of time series analysis and its use in the detection of anomalies in industrial networks. AR-X, ARIMA, SARIMA, Random Forest, Facebook Prophet and XGB Boost algorithms were used in the solution to create prediction models. In addition, the work includes the implementation of an algorithm for detecting anomalies from prediction models as well as solving the problem of high seasonal period in the case of the SARIMA algorithm. Through the conducted research, it was found that with the use of selected algorithms, it is possible to predict industrial traffic for the purpose of detection, within which up to 90% of attacks were detected. The work also provides a solution to a high seasonal period using partial time series. These results allow the experimental integration of prediction-based detection into real industrial networks.

Kľúčové slová

analýza časových radov, predikcie v časových radoch, detekcia anomálií v časových radoch, detekcia útokov v premávke, časový rad, AR-X, ARIMA, SARIMA, Random Forest, Facebook Prophet, XGB Boost

Keywords

time series analysis, time series forecasting, anomaly detection in time series, attack detection in network traffic, time series, AR-X, ARIMA, SARIMA, Random Forest, Facebook Prophet, XGB Boost

Citácia

BUDAI, Samuel. *Analýza časových řad.* Brno, 2023. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Ivana Burgetová, Ph.D.

Analýza časových řad

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pani doktorky Ing. Ivany Burgetovej Ph.D.. Uviedol som všetky literárne pramene, publikácie a ďalšie zdroje, s ktorých som čerpal.

.....
Samuel Budai
15. mája 2023

Poďakovanie

Veľmi rád by som úprimne poďakoval vedúcej mojej práce Ing. Ivane Burgetovej Ph.D. za jej ústretovosť a ochotu pomôcť mi nájsť riešenie vždy, keď som sa zasekol a nevedel sa pohnúť. Rovnako by som jej chcel veľmi pekne poďakovať za cenné rady a postrehy, ktoré ma častokrát naviedli správnym smerom a značne mi pomohli v rámci tvorby tejto práce.

Obsah

Kapitola 1

Úvod

Čas je vzácnou komoditou, ktorú ak raz stratíme, už nikdy nezískame späť. Vo svete priemyselných sietí, ktoré slúžia na riadenie a monitorovanie rôznych priemyselných procesov, hrá čas rozhodujúcu úlohu pri odhaľovaní a prevencii potenciálnych kybernetických útokov. Tieto siete zodpovedajú za pohodlný chod našich každodenných životov a riadia všetko od elektrární, cez dopravné systémy až po výrobné zariadenia. Ako si isto vieme predstaviť, tieto siete generujú obrovské množstvo údajov, ktoré poskytuje príležitosti pre detekciu anomálneho správania a predchádzanie kybernetickým útokom. Tieto útoky môžu spôsobiť podniku značné finančné straty, poškodenie dobrého mena a dokonca právnu zodpovednosť.

Schopnosť rýchlo a efektívne odhaliť a zabrániť kybernetickým útokom je v dnešnom svete poprepájanom internetom nevyhnutná. S rozmachom technológií, inteligentných strojov a zariadení sa priemyselné siete stávajú čoraz zložitejšími a náchylnejšími na kybernetické útoky, ktoré môžu mať ničivé následky. Preto je ich detekcia a prevencia rozhodujúca pre zabezpečenie hladkej a nepretržitej prevádzky týchto systémov.

Jedným z účinných nástrojov pre túto problematiku môže byť analýza časových radov, ktorá sa s narastajúcim výpočtovým výkonom stáva čoraz populárnejšou a možno ju využiť na identifikáciu vzorcov, ktoré môžu naznačovať potenciálny kybernetický útok. Jedná sa o štatistickú metódu, ktorá sa v dnešnej dobe vo veľkej miere využíva na modelovanie a predpovedanie sekvenčných údajov, akými sú sieťová prevádzka, systémové logy alebo bezpečnostné incidenty. Analýzou týchto údajov môžeme odhaliť zmeny a trendy ako napríklad nezvyčajné a neočakávané vzory v sieťovej prevádzke alebo náhle skoky, prípadne poklesy hodnôt snímačov, ktoré môžu byť indikátormi útokov. Táto práca je rozdelená do siedmych kapitol, kde prvou z nich je úvod. Druhá kapitola sa venuje základným pojmom analýzy časových radov a priblíženiu podstatných základov pre ich využitie, pojmom akými sú napríklad stacionarita časových radov, jeho významu, testom a samozrejme transformácií radov, ktoré toto kritérium nespĺňajú. Tretia kapitola sa venuje základným prístupom k analýze a predpovedaniu časových radov, ako aj priblíženiu základných modelov.

Kapitola 2

Základné pojmy analýzy časových radov

Táto kapitola sa venuje základným pojmom analýzy časových radov. V prvom rade priblíži samotný koncept časového radu a jeho základné typy. Vysvetlí, čo presne tento pojem znamená. Predstaví význam analýzy časových radov, ako aj to prečo je dôležitá pre pochopenie dát a predpovedanie budúcich hodnôt. Okrem toho kapitola predstaví základné problémy pri analýze a najčastejšie úpravy ktoré sa s časovými radmi vykonávajú.

V druhej časti sa kapitola venuje dôležitej vlastnosti časových radov zvanej stacionarita ako aj testom ktoré testujú prítomnosť tejto vlastnosti. Pre príklad priblíži štatistické testy ako ADF alebo KPSS.

2.1 Časový rad

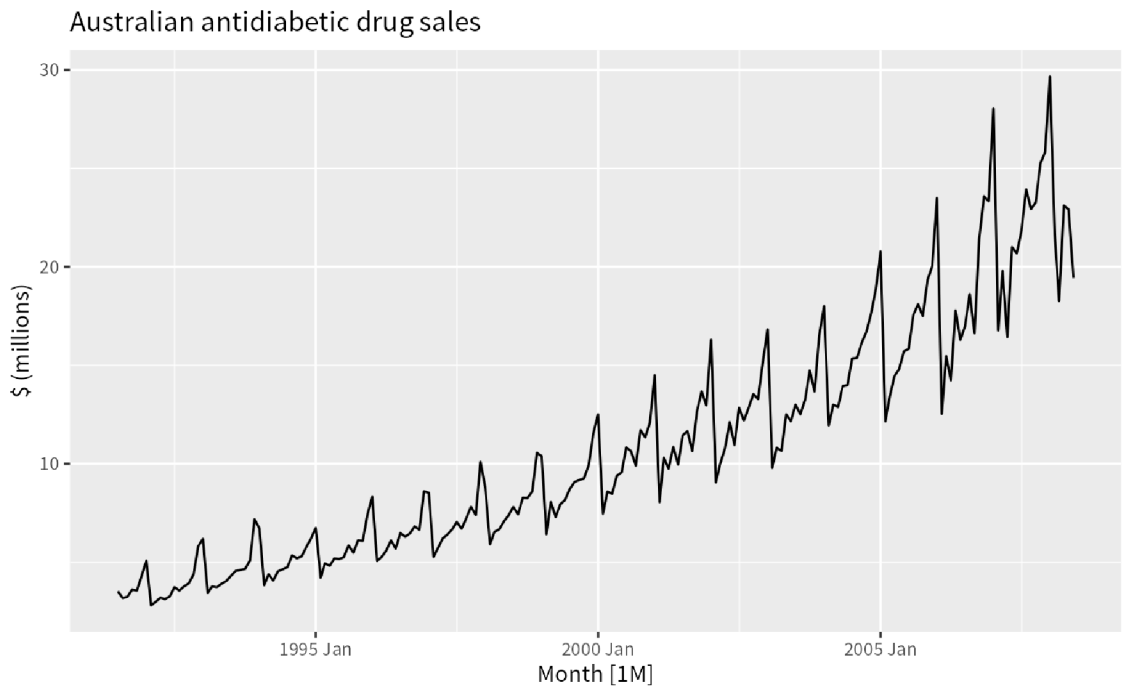
V štatistike sa pod pojmom časový rad rozumie postupnosť pozorovaní alebo meraní v určitých, zvyčajne po sebe nasledujúcich intervaloch. Dáta, ktoré vytvárajú časový rad, vznikajú ako chronologicky usporiadané pozorovania, pre ktoré je podstatné, že sú v čase usporiadané. Ako príklad si môžeme uviesť ceny akcií konkrétnej spoločnosti, seizmický záznam v geofyzike, rady najvyšších denných teplôt v meteorológii či vývoj cien v ekonómii. Príkladom bežného časového radu je aj záznam o predajoch istej kategórie lieku za určité obdobie, ktorý je možné vidieť na obrázku 2.1. Tieto intervaly sú spravidla rovnomerne rozmiestnené (ekvidištantné), a preto ich je možné zapísať nasledujúcim spôsobom:

$$y_1, y_2, \ldots, y_n \text{ alebo } y_t, t = 1, \ldots n$$

kde y znamená analyzovaný ukazovateľ, t je časová premenná s celkovým počtom pozorovaní n [9].

Pozorovania časových radov môžu byť zaznamenávané v rôznych frekvenciách v závislosti od skúmaného javu. Napríklad už spomínané ceny akcií sa môžu zaznamenávať každú sekundu alebo minútu, zatiaľ čo údaje o teplote môžu byť zaznamenávané každú hodinu, deň alebo každý mesiac.

Konkrétnejšie sa pojmom časový rad myslia stochastické procesy (sekvencie náhodných premenných), ktoré sú zaťažené neistotou a nie rady deterministické, ktorých správanie je možné popísať nejakým matematickým vzorcom. S využitím teórie náhodných procesov je možné povedať, že časový rad predstavuje konkrétnu realizáciu zodpovedajúcu stochastickému procesu [4, 9, 16].



Obr. 2.1: Mesačný predaj antidiabetík v Austrálii [13].

2.1.1 Druhy časových radov

Časové rady sa členia podľa charakteru ukazovateľa na [9]:

- Okamihové – udávajú hodnotu ukazovateľa k určitému okamihu t (napr. počet evidovaných uchádzačov)
- Intervalové – výsledná veľkosť sledovaného ukazovateľa závisí na dĺžke intervalu (obdobia), počas ktorého je ukazovateľ zaznamenávaný

Podľa druhu ukazovateľa sa rozlišujú časové rady obsahujúce [9]:

- Absolútne ukazovatele (očistené)
- Odvodené ukazovatele (súčtové, priemerové)

2.2 Význam analýzy časových radov

Pri vykonávaní analýzy časového radu je cieľom konštrukcia modelu, ktorý čo najlepšie popisuje, aké chovanie je možné od skúmaného časového radu očakávať. To umožní predovšetkým porozumieť mechanizmu, na základe ktorého sú sledované údaje generované a tiež či vykazujú pravidelné cykly alebo určitý trend. Pomocou takej konštrukcie je tiež možné zistiť, ako a na čom hodnoty časového radu závisia a naopak, čím pravdepodobne ovplyvňované nie sú. Model, ktorý sa pri tomto postupe získa, je možné následne využiť na predpovedanie budúcich hodnôt daného časového radu. Musíme ale pamätať na to, že hodnoty časového radu sú len náhodné pozorované veličiny a preto nami vytvorený model nie je schopný so stopercentnou určitosťou povedať presnú hodnotu v budúcnosti, ale vygeneruje iba jej odhad.

Z toho vyplýva, že znalosť modelu zodpovedá znalosti algoritmu, podľa ktorého by dáta generoval počítač, pričom vrámci algoritmu sú zapojené tiež generátory náhodných čísiel, dodávajúc celému procesu generovania náhodný charakter. Aj keď je možné typ generátorov a ich zapojenie do systému presne špecifikovať, na druhej strane nie je v žiadnom prípade možné stanoviť konkrétnu numerickú hodnotu produkovanú týmito generátormi v jednotlivých časových okamihoch. Znalosť modelu však umožňuje predpovedať budúci vývoj systému [4, 15].

2.3 Špecifické problémy analýzy časových radov

Táto kapitola stručne priblíži niektoré pomerne časté problémy, ktoré súvisia so špecifickým charakterom dát usporiadaných do časového radu.

- Problémy s dĺžkou časových radov:

  Pod pojmom dĺžka radu sa rozumieme príslušný počet n hodnôt, ktoré daný časový rad vytvárajú. Samozrejmosťou je, že s rastúcou dĺžkou časového radu sa zväčšuje množstvo informácií pre jej analýzu, avšak je potrebné upozorniť na to, že napr. zdvojnásobenie počtu meraní, ktoré je k dispozícii, nemusí nutne znamenať zdvojnásobenie množstva informácií obsiahnutých v týchto meraniach. S prihliadnutím na tento fakt, je potrebné si uvedomiť, že dĺžka radu nie je jednoznačnou mierou informácie obsiahnutej v rade.

  V tejto súvislosti sa často stretávajú dve protichodné tendencie. Na jednej strane niektoré metódy pre analýzu časových radov vyžadujú určitú minimálnu dĺžku radu (napr. Box-Jenkinsov prístup predpokladá minimálne 50 pozorovaní). Na strane druhej, u radov, ktoré sú pomerne dlhé, dochádza k nebezpečiu, že sa s priebehom času podstatne mení charakteristika modelu, ktorý tento rad generuje. Práve preto sa vybudovanie modelu s narastajúcou dĺžkou radu stáva náročnejšie, najmä v prípade modelov predpokladajúcich stabilné chovanie parametrov [4].

- Problémy s kalendárom:

  Tieto problémy sú len z malej časti spôsobené prírodou (napr. počet dní jedného slnečného roku nie je celé číslo), za väčšinu si však môže človek sám. Príkladom problémov, ktoré zavinil človek, sú rôzne dĺžky kalendárnych mesiacov, rozdielny počet víkendov v mesiaci (štyri alebo päť), rôzny počet pracovných dní v mesiaci alebo pohyblivé sviatky (napr. Veľká noc).

  Aj keď to na prvý pohľad nemusí byť očividné, nepravidelnosti, ktoré sa spájajú s kalendárom, môžu mať pomerne výrazné následky. Ak je napríklad sviatok umiestnený na začiatku mesiaca, potom sa vzhľadom k uzavretým obchodom počas sviatku zníži predaj potravín za aktuálny mesiac, ale zvýši sa predaj potravín za mesiac predchádzajúci. V prípade, že sa bude s takto ovplyvnenými dátami pracovať, je najprv potrebné ich štandardizovať. Zavádza sa napríklad tzv. "štandardný mesiac" s dĺžkou 30 dní, štandardný počet pracovných dní v mesiaci alebo sa pozorované údaje agregujú (napr. použitie kvartálnych dát namiesto mesačných) [4, 16].

- Problémy s voľbou časových bodov pozorovania:

  Diskrétne časové rady, tzn. rady, ktoré sú tvorené pozorovaniami v určitých nespojitých bodoch v čase, môžu vznikať tromi spôsobmi: sú priamo diskrétne svojou povahou (napr. úroda obilia za jednotlivé roky), alebo vznikajú diskretizáciou spojitého

časového radu (napr. teplota v danú dennú dobu na určitom danom mieste), alebo agregáciou (akumuláciou) hodnôt za dané časové obdobie (napr. denné množstvo zrážok, alebo počet nalietaných kilometrov za mesiac jednou leteckou spoločnosťou; namiesto akumulácie hodnôt sa častokrát robí ich priemer).

Je samozrejmé, že v niektorých prípadoch človek vykonávajúci analýzu nemá žiadny vplyv na štruktúru, interval či voľbu časových bodov pozorovania. Ak však ale túto možnosť má, mal by jej venovať určité množstvo času s cieľom nájsť kompromis medzi častokrát protichodnými požiadavkami. Veľká hustota časových bodov pozorovania síce umožňuje dobre vystihnúť charakteristické rysy, ale z hľadiska numerickej jednoduchosti a časovej náročnosti výpočtu pri analýze časových radov, je nežiadúce príliš "zahusťovať" počet pozorovaní, pretože by mohli nastať problémy pri výpočtoch. Na druhú stranu pozorovania nesmú byť ani príliš riedke, nakoľko by analytikovi mohli uniknúť niektoré podstatné a charakteristické črty daného radu. Ak sú skúmaným faktorom napr. zmeny časového radu v priebehu rokov, je potrebné mať k dispozícii aspoň niekoľko pozorovaní behom každého roku a najlepšie v tých istých časových intervaloch. Obvykle je primárnou snahou pracovať s pozorovaniami, ktorých dĺžky intervalov sú ekvidištantné (tzn. rovnako vzdialené) [4].

- Problémy s nezrovnalosťou jednotlivých meraní:

  Problémy s nezrovnalosťou meraní súvisia s výberovou vzorkou a zároveň reprezentatívnosťou tejto vzorky aj z hľadiska časového vývoja (napr. dáta v prvom mesiaci sú zo 64 podnikov, no v druhom mesiaci len zo 47 podnikov) [9]. Takýto problém by sa na základe charakteru dát dal vyriešiť dvomi spôsobmi, buďto záznamy s chýbajúcimi hodnotami odstránime. alebo sa pokúsime chýbajúce hodnoty odhadnúť. Hodnoty pridáme odhadom však môžu skresľovať vytvorený model. Základné úpravy časových radov budú popísané v nasledujúcej kapitole.

2.4 Základné úpravy časových radov

Pre prácu s časovými radmi je dôležité uviesť a vysvetliť najčastejšie transformácie a úpravy časových radov.

2.4.1 Doplnenie chýbajúcich hodnôt

Môže sa stať, že v časovom rade budú niektoré hodnoty pozorovaní chýbať, a zvyčajne je potrebné tieto hodnoty pred analýzou doplniť. Takéto údaje sa samozrejme nedajú považovať za plnohodnotné a ich prítomnosť znižuje vierohodnosť a spoľahlivosť analýzy. Ako príklad je možné uviesť niekoľko prístupov snažiacich sa o odstránenie daného problému, z ktorých sa volí na základe účelu analýzy:

- Najjednoduchším prístupom je doplniť chýbajúce hodnoty nulami. No tento prístup by sa mal používať len v prípade, ak o časovom rade nevieme vôbec nič, alebo len to, že jeho priemerný člen by mal byť nulový.

- Nahradenie chýbajúcich hodnôt aritmetickým priemerom alebo mediánom, kde sa tento problém dá riešiť v globálnej mierke a s tým, že sa berie ohľad na charakteristiku celej dátovej sady, alebo v mierke lokálnej a dáta sa dopĺňajú len s využitím okolitých známych bodov.

- Pre rady, ktoré vykazujú výraznú zotrvačnosť sa hodí nahradenie chýbajúcich hodnôt lineárnou interpoláciou susedných bodov.
- Nahradenie chýbajúcich hodnôt odhadom založeným na známom, alebo odhadnutom modely chovania procesu [9].

2.4.2 Nelineárna transformácia mierky

Táto metóda sa používa predovšetkým pre zmiernenie či potlačenie nestacionarity časového radu v prípade, ak napríklad, s narastajúcimi hodnotami radu rastie aj rozptyl jeho členov. V takom prípade môžeme využiť logaritmovanie alebo odmocnenie radu, ktoré môže tento problém potlačiť. Po vykonaní analýzy sa k pôvodnej mierke dostaneme spätnou transformáciou, napr. ak sa časová radu odmocnila, spätná transformácia sa dostane umocnením. V prípade využitia logaritmovania, sa tento proces vráti transformáciou pomocou exponenciálnej funkcie [9].

2.4.3 Časový posun

Ako už názov napovedá, jedná sa o posun pôvodného časového radu dopredu alebo dozadu oproti pôvodnému časovému radu. Táto metóda pozostáva z vytvorenia nového časového radu, ktorý je vzhľadom k pôvodnému časovému radu oneskorený, alebo tento časový rad predbieha, ale inak je s ním totožný. Je dôležité podotknúť, že v novo vytvorenom rade na začiatku, resp. na konci (podľa smeru, v ktorom sa posun vykonával) chýba toľko hodnôt, o koľko krokov sa vykonal posun [9].

2.4.4 Sezónna diferenciácia

Sezónna diferencia časového radu odkazuje na proces odčítania pozorovaní v časovom rade a k nemu korešpondujúcich pozorovaní s predchádzajúcej sezóny. Pre údaje s intervalom jedného mesiaca (s ročným sezónnym cyklom) sa pojmom sezónna diferencia prvého rádu rozumie rozdiel údajov z tohtoročného mesiaca s k nemu príslušným mesiacom minulého roku. Diferencia vyjadruje veľkosť zmeny, ku ktorej došlo medzi dvoma časovými úsekmi merania. Ak diferencia vyjde záporná znamená to, že časový rad klesá. Naopak ak je výsledná diferencia kladná, rad v danom čase rastie. Pomocou diferenciácie je možné pozorované dáta zbaviť lineárneho trendu a podobne pomocou sezónnych diferencií je možné dáta zbaviť sezónnych vplyvov [16, 18].

2.4.5 Kumulatívny súčet

Hodnota kumulatívneho súčtu sa získa tak, že sa postupne sčítajú všetky hodnoty radu od jeho počiatku až po určitý okamih v čase (koniec radu). Jeho výsledná hodnota teda udáva súčet všetkých hodnôt, ktoré sa vyskytli v rade v danom okamihu. Ak postupne vykonáme aplikáciu diferenciácie a kumulatívneho súčtu, získame pôvodný rad oneskorený o jeden časový interval, ktorý môže byť zväčšený alebo zmenšený o určitú konštantu [9].

V tejto kapitole je dôležité spomenúť, že zaujímavým časovým radom je rad, ktorý vznikne kumulatívnym súčtom bieleho šumu a nazýva sa náhodná prechádzka, alebo aj prechádzka „opitého námorníka" ktorej príklad grafického vyobrazenia je možné vidieť na obrázku 2.2. Toto pomenovanie dostala vďaka svojej vlastnosti - nemožnosti predvídať, či sa táto funkcia obráti hore alebo dole. Podľa tejto zákonitosti by sa mali riadiť napr. ceny akcií na burze [9].

Obr. 2.2: Príklad procesu náhodnej prechádzky [18]

2.4.6 Vyhladzovanie časových radov

V prípade, ak je nejaká veličina (napr. časový rad) meraná v príliš krátkych časových intervaloch, môže sa stať, že po sebe nasledujúce hodnoty budú odlišné len veľmi malou odchýlkou od predchádzajúcej hodnoty. Tieto odchýlky sú obvykle náhodného charakteru a môžu byť spôsobené rôznymi faktormi, ako sú šumy v meracích zariadeniach alebo zmeny okolitého prostredia.

V prípade, že tieto náhodné odchýlky majú strednú hodnotu nula (tzn. táto chyba v priemere hodnotu zväčší a zmenší rovnako často) a ak sú tieto odchýlky na sebe nezávislé (tzn. neexistuje medzi nimi korelácia), je možné predpokladať, že spriemerovaním niekoľkých po sebe idúcich hodnôt, majú tieto náhodné odchýlky tendenciu vyrušiť sa, a tak vynikne skutočná hodnota sledovaného procesu. Tento proces sa opakuje pre každú nasledujúcu hodnotu časového radu, a tým sa vytvára nový vyhladený časový rad [9].

Príkladom metód vyhladzovania časových radov môžu byť nasledujúce prístupy:

- Stredové kĺzavé priemery: hodnotu nahrádzame aritmetickým priemerom samej seba a najbližších predchádzajúcich a najbližších nasledujúcich pozorovaní, ležiacich najďalej do určenej časovej vzdialenosti (tzn. prístup posúvajúceho sa okna)
- Kĺzavé priemery s predchádzajúcich hodnôt: hodnotu nahrádzame aritmetickým priemerom samej seba a zväčša všetkých predchádzajúcich pozorovaní
- Kĺzavé mediány: hodnotu nahrádzame mediánom samej seba a najbližších pozorovaní v určitom časovom okne [9].

2.5 Stacionarita časových radov

Pri analýze časových radov je stacionarita časového radu dôležitým pojmom, ktorý odkazuje na štatistické vlastnosti radu, ktoré sa v čase nemenia – sú teda konštantné. Stacionarita je nevyhnutná pre mnoho modelov a techník analýzy časových radov, pretože zabezpečuje, že údaje majú predvídateľné a konzistentné správanie, ktoré možno presne modelovať a predpovedať.

2.5.1 Priemer, Variancia, Autokorelácia

Predtým, ako bude vysvetlený pojem stacionarita, je dôležité definovať pár základných termínov, ktoré s týmto pojmom úzko súvisia.

- Priemer - priemerná hodnota pozorovaní vo všetkých časových bodoch. Predstavuje strednú hodnotu dát a poskytuje pohľad na typickú úroveň radu. Konštantný priemer v priebehu času naznačuje, že dáta nie sú ovplyvnené žiadnymi dlhodobými trendmi alebo systematickými vzormi, ktoré by ovplyvnili celkovú úroveň časového radu [24].

- Variancia (rozptyl) – sa používa v štatistike pre opísanie miery variability (rozptylu) medzi hodnotami dátovej sady a jej priemerom. Poskytuje informácie o miere náhodnosti (volatility) v údajoch. Čím väčšia je hodnota variancie, tým väčšia je vzdialenosť medzi číslami v dátovej sade a priemerom. Naopak, menšia hodnota variancie znamená, že čísla v sade sú blízko k priemeru [8].

- Autokorelácia - je miera podobnosti medzi daným časovým radom a posunutou verziou samého seba cez postupné časové intervaly. Zjednodušene povedané, autokorelácia sa snaží merať vzťah medzi súčasnou hodnotou premennej a akýmikoľvek minulými hodnotami, ktoré má k dispozícii.

  Odráža tendenciu časového radu prejavovať vzory alebo závislosti v čase. Autokorelácia môže byť kladná, čo naznačuje, že pozorovania majú tendenciu podobať sa v predchádzajúcich časových bodoch, alebo záporná, čo naznačuje, že pozorovania majú tendenciu byť odlišné od tých v predchádzajúcich časových bodoch. Konštantná autokorelácia v čase naznačuje, že vzor závislosti medzi pozorovaniami zostáva v čase rovnaký [7].

2.5.2 Stacionarita

Jedná sa o veľmi dôležitú charakteristiku časových radov. Hovoríme, že časový rad je stacionárny, ak sa jeho štatistické vlastnosti nemenia v čase, teda jeho vlastnosti nie sú závislé od času, v ktorom je daný časový rad pozorovaný. S určitosťou je možné povedať, že časové rady s trendom alebo sezónnou zložkou stacionárne nie sú - trend a sezónna zložka ovplyvnia hodnotu radu v rôznych časových úsekoch. Nie vždy je však stacionarita na prvý pohľad jasná a určiť stacionaritu môže mnohým robiť problémy - rad s cyklickým správaním (ale bez trendu alebo sezónnej zložky) môže byť stacionárnym časovým radom. Je to spôsobené faktom, že cykly nemajú pevne viazanú dĺžku a frekvenciu opakovania, takže pred pozorovaním nie je isté, kde a kedy sa vyskytnú vrcholy a minimá cyklov [12].

Ďalšie zdroje uvádzajú rozdelenie stacionarity na striktnú a slabú. Striktná stacionarita predpokladá, že správanie príslušného náhodného procesu, tzn. jeho rozdelenie, je invariantné voči časovým posunom. Na rozdiel od toho slabá stacionarita je menej obmedzujúca; požaduje, aby časový rad spĺňal kritéria konštantnosti priemeru, variancie a auto-korelácie v čase.

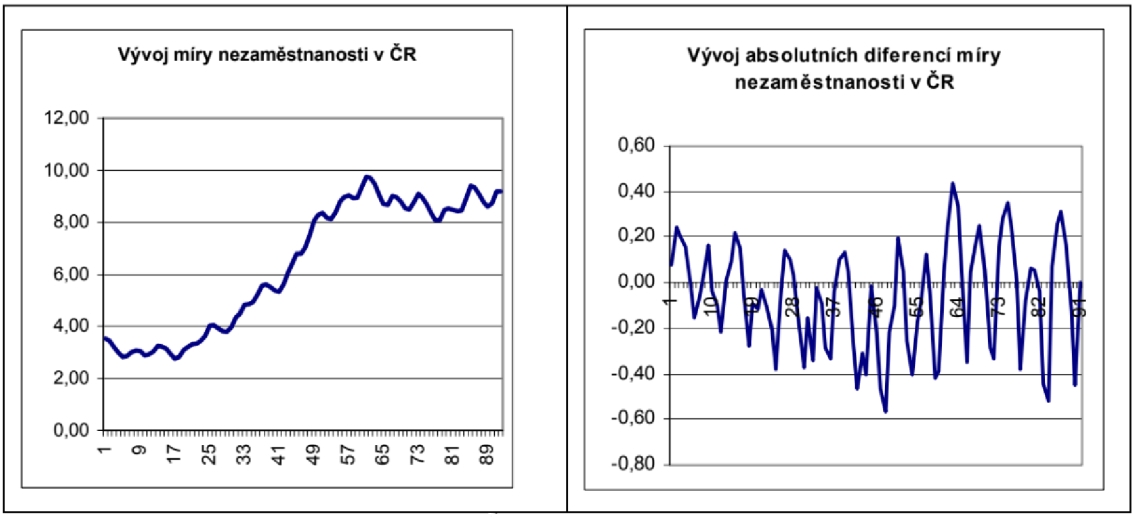
Pod konštantným priemerom časového radu myslíme dáta, ktoré by nemali vykazovať žiadny dlhodobý trend, alebo obsahovať systematické vzory, ktoré ovplyvňujú časový rad. Konštantná variancia udáva, že by časový rad nemal vykazovať žiadne konštantné zmeny v úrovni volatility alebo náhodnosti v čase a konštantná autokorelácia zjednodušene hovorí, že vzor závislosti medzi pozorovaniami zostáva v priebehu času rovnaký. Túto vlastnosť môžeme vyjadriť vzťahom, kde pre kovariancie platí

$$cov(Y_t, Y_s) = cov(Y_{t+h}, Y_{s+h})$$

a to pre ľubovoľné h [16]. Uvedený vzťah predstavuje požiadavku, aby závislosť medzi dvoma ľubovoľnými pozorovaniami závisela len na ich časovej vzdialenosti a na ich časovom umiestnení v rade [9, 16, 23].

V tejto práci sa bude ďalej pod pojmom stacionarita rozumieť slabá stacionarita.

Správanie časového radu môže teda zo štatistického hľadiska podliehať buď zmenám v priemere alebo variabilite (rad nestacionárny), alebo byť stále rovnaké (rad stacionárny). Zjednodušene by sa dalo povedať, že u stacionárneho radu nie je takmer možné na základe zistených štatistických parametrov, ako sú aritmetický priemer hodnôt alebo ich rozptyl, odlíšiť jeden úsek radu od druhého. Nestacionárny rad naopak vykazuje zmeny v správaní: napríklad aritmetický priemer hodnôt na začiatku radu je signifikantnejšie rozdielny ako priemer členov na konci (o takomto rade hovoríme, že vykazuje trend). Stacionarita je podstatným predpokladom niektorých typov analýz, preto je potrebné stacionaritu testovať a v prípade nesplnenia tejto podmienky, je následne potrebné radu vhodným spôsobom transformovať s cieľom zabezpečenia stacionarity [9].



Obr. 2.3: Vývoj mesačnej miery nezamestnanosti v ČR od roku 1995 do polovice roku 2002 [9]

Na obrázku 2.3 sú vyobrazené dva grafy, na ktorých je možné pozorovať priebeh typického nestacionárneho časového radu, ukazujúci rastúci trend, sezónne vplyvy v priebehu každého roka a s časom rastúci rozptyl (sezónne odchýlky od priemeru sa stále zväčšujú). Takýto rad nevykazuje žiadnu časovú zmenu parametrov, pretože jeho všeobecný člen nezávisí ani od času, ani od predchádzajúcich členov radu [9].

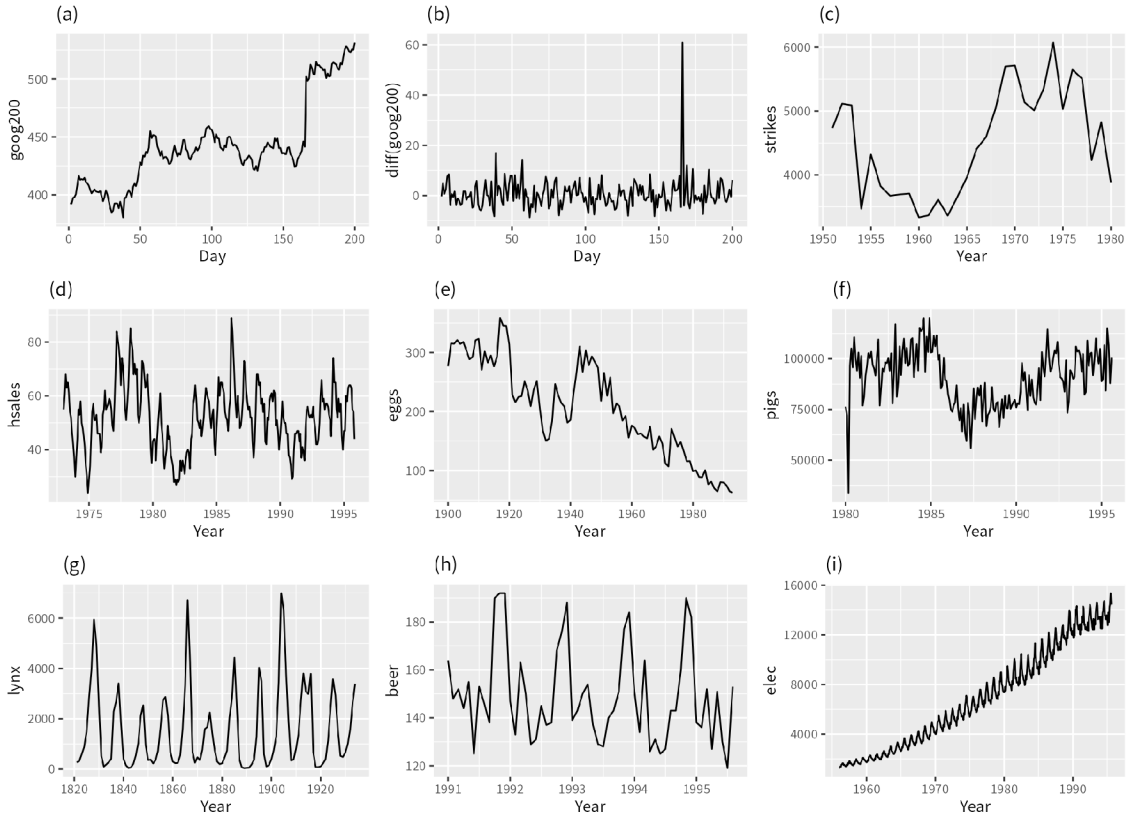
2.5.3 Testovanie stacionarity

V analýze časových radov je veľmi dôležitá schopnosť určiť stacionaritu tohoto radu. V praxi to však zvyčajne znamená, že sa analytik nesnaží rozhodovať sa medzi dvomi striktnými možnosťami, ale určiť, s čo najvyššou pravdepodobnosťou, pomocou rôznych prístupov, či je časový rad tvorený stacionárnym procesom [1].

2.5.3.1 Vizuálny test

Najjednoduchšia metóda pre detekciu stacionárnosti spočíva vo vizuálnej kontrole časového radu. Zahŕňa vykreslenie dát alebo ich funkcií a vizuálne preskúmanie, či graf vykazuje nejaké známe vlastnosti stacionárnych (príp. nestacionárnych dát). Ako už bolo spomenuté, ak je na prvý pohľad rozoznateľné, že sa priemer alebo rozptyl údajov zdajú byť v priebehu času konštantné, potom je rad pravdepodobne stacionárny, avšak, ak sa priemer a rozptyl v čase viditeľne menia, potom je rad pravdepodobne nestacionárny. Podobne, ak je v grafe jasne prítomný trend alebo sezónnosť, rad bude pravdepodobne nestacionárny.

Hoci môže byť vizuálna kontrola rýchlym a jednoduchým spôsobom na posúdenie stacionárnosti, tieto vlastnosti nemusia byť vždy z grafu na prvý pohľad zreteľné. S prihliadnutím na tento fakt sa nedá povedať, že by táto metóda bola príliš spoľahlivá, preto by mala slúžiť skôr na získanie prvého obrazu o dátach a pre konečné závery by mali byť použité nejaké z existujúcich spoľahlivejších štatistických metód [1, 12].



Obr. 2.4: Ktoré z týchto časových radov sú stacionárne? (a) Cena akcií Google počas 200 po sebe nasledujúcich dní; (b) Denná zmena ceny akcií Google počas 200 po sebe nasledujúcich dní; (c) Ročný počet štrajkov v USA; (d) Mesačný predaj nových rodinných domov v USA; (e) Ročná cena dvanástich vajec v USA (v dolároch); (f) Mesačný celkový počet zabitých ošípaných vo Victorii, Austrália; (g) Ročný celkový počet uviaznutých rysov v oblasti rieky McKenzie v severozápadnej Kanade; (h) Mesačná produkcia austrálskeho piva; (i) Mesačná produkcia austrálskej elektriny [12]

.

Uvažujeme deväť časových radov vyobrazených na obrázku 2.4. Skúsime si rozobrať, ktorý z týchto radov je stacionárny.

Na prvý pohľad je zrejmé, že sezónna zložka vylúči rady (d), (h) a (i). Trendy a meniace sa úrovne vylúčia rady (a), (c), (e), (f) a (i). Narastajúca variancia v čase tiež vylúči rad (i). Tým pádom ako stacionárne rady zostávajú rady (b) a (g).

Niekomu by sa mohlo zdať, že silná cyklická zložka v rade (g) by ju mohla ovplyvniť natoľko, že by sa jednalo o nestacionárnu radu. Avšak tieto cykly sú aperiodické - vznikajú v prípade, keď sa populácia rysov príliš zväčší v prihliadnutí na dostupnú potravu. Ak tento prípad nastane, rysy sa prestanú rozmnožovať a populácia klesne na nižšie počty, následná regenerácia ich zdrojov potravy umožní populácii opäť rásť. Z dlhodobého hľadiska nie je načasovanie týchto cyklov predvídateľné. Séria je preto stacionárna [12].

Ďalšie vizuálne metódy na detekciu stacionarity zahŕňajú skúmanie auto-korelačnej funkcie (ACF) a čiastočnej auto-korelačnej funkcie (PACF) daného časového radu. Tieto grafy môžu pomôcť identifikovať, či je séria stacionárna, má trend alebo vykazuje sezónnosť [4].

2.5.3.2 Upravený test Dickeyho a Fullera

Pred vysvetlením nasledujúcich dvoch štatistických testov, je nutné spomenúť pojem jednotkový koreň a ako sa vzťahuje k časovým radom a stacionarite. Jednotkový koreň odkazuje na prítomnosť stochastického trendu v časovom rade, čo znamená, že rad nemá stabilný priemer alebo rozptyl. Ak časový rad obsahuje jednotkový koreň, hovorí sa, že je nestacionárny, čo znamená, že jeho štatistické vlastnosti sa v čase menia a je veľmi zložité vytvárať zmysluplné predpovede alebo závery z takýchto dát, pretože v takomto modeli môže dochádzať k výkyvom a otrasom. Pojem jednotkový koreň je štatistickým pomenovaním pre nestacionaritu, ktorá bola popísaná v predchádzajúcej kapitole. Prítomnosť jednotkového koreňa sa častokrát popisuje tak, že v časovom rade existuje, ak v nasledujúcej rovnici hodnota alfa je rovná jednej:

$$Y_t = \alpha Y_{t-1} + \beta X_e + \epsilon$$

kde Yt je hodnota časového radu v čase t, Y_{t-1} je oneskorená hodnota časového radu v čase $t-1$, α je koeficient oneskoreného časového radu, ktorý udáva prítomnosť jednotkového koreňa, X_e predstavuje prediktor alebo súbor prediktorov, ktoré potencionálne súvisia s časovým radom $Y(t)$. Písmeno X označuje maticu prediktorov a písmeno náhodnú zložku (tzv. error term), β meria vplyv exogénnej premennej X_e na aktuálnu hodnotu časového radu. A $e(t)$ je náhodná zložka [21].

Dickey Fullerov test je štatistický test hypotéz, ktorý meria množstvo stochasticity (náhodných procesov) v modeli časového radu (testuje prítomnosť jednotkového koreňa). Tento test je založený na jednoduchej lineárnej regresii, ktorá skúma, či prvá diferencia časového radu je stacionárna. Tento test nám vlastne vytvára t-štatistiku, ktorá je následne porovnávaná s vopred určenou kritickou hodnotou. Ak sa hodnota nachádza pod kritickou hodnotou, umožňuje to zamietnuť nulovú hypotézu a prijať alternatívu. V prípade, že daná hodnota presiahla kritickú hodnotu, nulovú hypotézu zamietnuť nedokážeme. Hypotézy:

$$H_0 = \text{Obsahuje proces koreňovej jednotky}$$
$$H_a = \text{Neobsahuje proces koreňovej jednotky}$$

Upravený test Dickeyho a Fullera je upravenou verziou Dickeyho Fullerovho pôvodného testu, ktorý v modeli zahŕňa vyšší rád regresných procesov; nulová hypotéza sa oproti

pôvodnému testu nijako nemení a teda ostáva rovnaká. To z neho robí mocnejšiu a flexibilnejšiu verziu pôvodného testu a umožňuje, aby bol použitý vo väčšom spektre kontextu [10, 17, 19, 20, 21].

2.5.3.3 Kwiatkowski Phillips Schmidt Shin test

KPSS (Kwiatkowski-Phillips-Schmidt-Shin) je štatistický test používaný na určenie stacionarity časového radu. Na rozdiel od Dickeyho-Fullerovho testu, ktorý testuje prítomnosť jednotkového koreňa (štatistický termín pre nestacionarita), test KPSS priamo testuje stacionaritu.

Test KPSS je založený na nulovej hypotéze, že časový rad je stacionárny a na alternatívnej hypotéze, že tento rad stacionárny nie je. Štatistika testu sa vypočíta na základe rozptylu časového radu a nulová hypotéza sa zamietne, ak je štatistika testu nad kritickou hodnotou.

Test KPSS sa často používa v spojení s testom Dickey-Fuller, pričom tieto dva testy poskytujú doplnkové informácie o stacionárnosti časového radu. Dickey-Fullerov test dokáže identifikovať, či séria má jednotkový koreň, zatiaľ čo test KPSS môže určiť, či je séria stacionárna alebo nie.

V praxi výber medzi testom Dickey-Fuller a KPSS závisí od špecifických charakteristík analyzovaného časového radu a skúmanej otázky. Ak je cieľom určiť, či je séria stacionárna alebo nestacionárna, môže byť vhodnejší test KPSS. Ak je však cieľom identifikovať prítomnosť koreňovej jednotky, môže byť užitočnejší Dickey-Fullerov test [17, 22].

Kapitola 3

Základné prístupy k analýze a predpovedi časových radov

Existuje mnoho prístupov k tomu, akým spôsobom je možné časový rad analyzovať. Voľba prístupu k analýze závisí na celej rade faktorov: typ sledovaného časového radu, nakoľko niektoré metódy sú vhodné len pre časové rady vymedzeného typu, účel analýzy, ktorým je zvyčajne rozpoznanie mechanizmu generovania hodnôt časového radu a predpovedanie jeho budúceho vývoja, skúsenosti, štatistika, ale aj dostupnosť výpočtovej techniky a software. Štyrmi najčastejšie používanými metódami, ktoré budú v tejto kapitole priblížené, sú dekompozícia časového radu, Box-Jenkinsonova metodológia, lineárne kauzálne modely a spektrálna analýza časových radov [4].

3.1 Dekompozícia časového radu

Princíp tejto metódy je pomerne jednoduchý. Časový rad je môžné rozložiť na štyri základné zložky a tými sú trend (Tr), sezónna zložka (Sz), cyklická zložka (C) a náhodná (reziduálna) zložka (ϵ). Pričom dôraz sa kladie na systémové zložky a jednotlivé pozorovania sa zvyčajne berú ako nekorelované (vzájomne nezávislé) [6, 16].

Motivácia pre dekompozíciu časového radu pramení s domnienky, že v jednotlivých zložkách rozkladu sa podarí jednoduchšie, než v pôvodnom nerozloženom rade, identifikovať pravidelné chovanie radu [4].

1. Trend odráža dlhodobý vzorec alebo dlhodobé zmeny v priemernom chovaní časového radu. Môžeme ho pozorovať v prípade, ak dochádza k dlhodobému rastu alebo poklesu (príp. dlhodobá konštantná úroveň) hodnôt dát pozorovaní a je možné si predstaviť, že trendová zložka vzniká v dôsledku pôsobenia síl, ktoré pôsobia systematicky v rovnakom smere . Trendy nemusia byť vždy lineárne, v praxi sa častokrát stretávame s trendmi nelineárnymi. Tieto trendy môžu byť ovplyvnené faktormi, ako sú technologický pokrok, demografické zmeny alebo ekonomické podmienky [6, 13].

2. Sezónna zložka predstavuje periodické zmeny v časovom rade, ktoré sa odohrávajú v priebehu jednej sezóny (napríklad jeden kalendárny rok) a každú ďalšiu sezónu sa opakujú. Tieto zmeny spravidla súvisia so striedaním ročných období. Dôležité je podotknúť, že sezónnosť má vždy fixnú a vopred známu periódu [13, 16].

3. Cyklická zložka je najspornejšou zložkou časového radu. Jedná sa o periodickú zložku, ktorá sa v mnohých literatúrach popisuje ako fluktuácia okolo trendu, pri ktorej sa

pravidelne striedajú fázy rastu s fázami poklesu. Dĺžka jednotlivých cyklov a intenzita jednotlivých fáz sa pritom môžu v priebehu času meniť. Cyklická zložka môže byť dôsledkom evidentných vonkajších vplyvov, avšak príčiny vedúce k vzniku cyklickej zložky sa spravidla identifikujú len veľmi ťažko [6, 16].

Mnoho ľudí si mýli cyklické správanie so sezónnym správaním, ale treba podotknúť, že tieto dve zložky sú úplne odlišné. Ak fluktuácie nemajú pevne stanovenú frekvenciu, potom sa jedná o cyklické správanie; ak sa však frekvencia v čase nemení a súvisí s niektorým aspektom kalendára, potom sa jedná o správanie sezónne. Vo všeobecnosti je priemerná dĺžka cyklov zvyčajne dlhšia ako dĺžka sezónneho modelu a veľkosť cyklov má tendenciu byť variabilnejšia ako veľkosť sezónnych modelov [13].

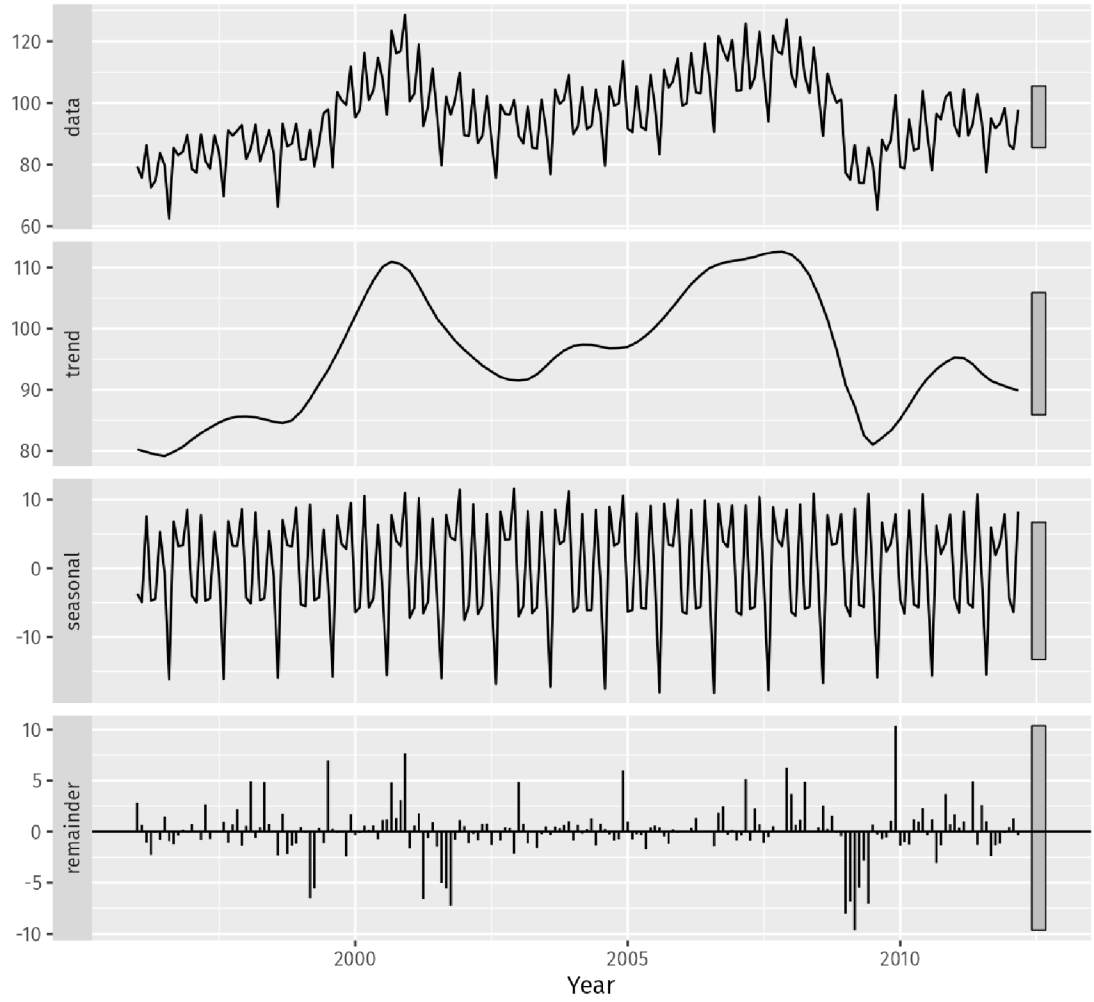
4. Náhodná (reziduálna) zložka, tiež známa ako nepravidelnosť alebo šum, predstavuje nepredvídateľné výkyvy (náhodné fluktuácie) v časovom rade, ktoré nemajú systematický (rozpoznateľný) charakter, nemožno ich vysvetliť trendom, sezónnosťou alebo cyklickosťou. Preto sa taktiež nepočíta medzi predchádzajúce tzv. systematické zložky časového radu. Zahŕňa aj faktory, akými sú chyby merania údajov, náhodné otrasy, chyby v štatistickom spracovaní dát alebo neočakávané udalosti. Často sa predpokladá, že má náhodná zložka charakter bieleho šumu, tzn. že je tvorená hodnotami nezávislých náhodných veličín s nulovou strednou hodnotou a nejakým konštantným rozptylom, dokonca sa niekedy predpokladá, že sa jedná o biely šum s normálnou distribúciou [16].

Poznáme dva typy dekompozície časového radu:

a) Aditívnu dekompozíciu v tvare:

$$Y_t = Tr_t + Sz_t + C_t + \varepsilon_t,$$

kde pri aditívnom rozklade sú všetky jednotlivé zložky uvažované vo svojich skutočných absolútnych hodnotách a po dekompozícii sú jednotlivé zložky v rovnakých merných jednotkách ako pôvodný časový rad (Y_t) [16]. Príklad takejto dekompozície môžeme graficky vidieť na obrázku 3.1. Používa sa v prípade, že variabilita hodnôt v čase je približne konštantná [9, 16].



Obr. 3.1: Aditívna dekompozícia časového radu [12]

b) Multiplikatívna v tvare:

$$Y_t = Tr_t Sz_t C_t \varepsilon_t,$$

kde po dekompozícii je v rovnakých merných jednotkách ako pôvodný časový rad (Y_t) iba trendová zložka, ale ostatné zložky sú uvažované v relatívnych hodnotách voči trendu a jedná sa teda o bezrozmerné veličiny [16]. Tento spôsob dekompozície sa používa v prípade, že variabilita časového radu rastie v čase alebo sa v čase mení [9, 16].

3.2 Box-Jenkinsonova metodológia

Z podkapitoly o dekompozícii by malo byť zrejmé, že dekompozičné metódy kladú dôraz predovšetkým na prácu so systematickými zložkami časového radu a že jednotlivé pozorovania sa obvykle berú ako nekorelované. Na rozdiel od toho, Box-Jenkinsova metodológia, ako základný prvok konštrukcie modelu časového radu, berie v úvahu reziduálnu (náhodnú)

zložku, ktorá môže byť tvorená korelovanými (závislými) náhodnými veličinami. Táto metodológia teda nielenže umožňuje spracovávať časové rady s navzájom závislými pozorovaniami, ale ťažisko ich postupov spočíva práve vo vyšetrovaní týchto závislostí, v tzv. korelačnej analýze. Rovnako táto metodológia umožňuje modelovať aj časové rady s výrazným trendovým a/alebo sezónnym charakterom, pričom trendová aj sezónna zložka môžu byť na rozdiel od dekompozičných metód modelované stochasticky. Pri používaní tohto prístupu sa zvyčajne predpokladá, že časový rad je slabo stacionárny. Box-Jenkinsonove modely sú výrazne flexibilnejšie ako modely dekompozičné, kde pod pojmom flexibilita rozumieme rýchlu schopnosť modelu adaptovať sa na zmeny v priebehu časového radu. Môže byť nápomocné si uvedomiť, že táto flexibilita modelu je dosiahnutá na základe tvorby modelu priamo z dát -– obvykle nie je možné navrhnúť tento model hypoteticky [4, 6, 9, 11, 16].

Akýkoľvek lineárny model môžeme obecne zapísať v tvare [6]:

$$y = G.u + H.e$$

kde y je výstupná veličina, G je prenosová funkcia, ktorá vyjadruje závislosť výstupu na vstupe (dynamické vlastnosti systému), u predstavuje vstupnú veličinu, H je tzv. "noise model" (akým spôsobom pôsobí rušenie na výstup zo štandardizovaného zdroja rušenia e(t)) a e predstavuje biely šum, teda skutočnosť, že s rovnakým vstupom dostaneme rôzny výstup.

V priebehu analýzy časového radu s použitím tejto metodológie je postup možné rozdeliť na tri primárne fázy:

1. Identifikácia modelu - táto fáza zahŕňa výber vhodnej štruktúry modelu, ktorá dokáže zachytiť vzory a charakteristiky v časovom rade. Prvým krokom pri vytváraní Box-Jenkinsovho modelu je kontrola splnenia podmienok stacionarity daného časového radu a zároveň určenie prítomnosti významnej sezónnej zložky, na ktorú bude potrebné prihliadať pri modelovaní. Akonáhle boli otázky stacionarity a sezónnosti vyriešené, poslednou časťou tejto fázy je určenie rádu auto-regresívnych podmienok a podmienok kĺzavého priemeru na základe auto-korelačnej a parciálne auto-korelačnej funkcie. Po identifikácii štruktúry modelu nasleduje fáza odhadu parametrov [23].

2. Odhad parametrov modelu - odhad parametrov pre Box-Jenkinsove modely je dosť komplikovaný nelineárny odhadovací problém. Z tohto dôvodu by mal byť odhad parametrov zverený kvalitnému softvérovému programu, ktorý dokáže modely Box-Jenkins prispôsobiť. Našťastie, mnoho komerčných štatistických softvérových programov teraz dokáže prispôsobiť modely Box-Jenkins. Základné prístupy, ktoré sa používajú pre odhadovanie parametrov sú: nelineárna metóda najmenších štvorcov a odhad maximálnej vierohodnosti, ktorá je zvyčajne uprednostňovaná [23].

3. Overovanie modelu - overovanie Box-Jenkinsovho modelu je principiálne podobné validácii modelu pre nelineárne odhady najmenších štvorcov. To znamená, že sa predpokladá, že chybová podmienka spĺňa predpoklady pre stacionárny jednorozmerný proces. Rezíduá by mali vykazovať vlastnosti bieleho šumu, čo znamená, že by mali byť náhodné, bez vzorov alebo trendov, alebo by alternatívne mali byť nezávislé, ak sú ich distribúcie normálne. Okrem toho by mali mať fixnú distribúciu s konštantným priemerom a rozptylom, čo naznačuje, že nie sú systematicky zaujaté alebo ovplyvnené externými faktormi. Ak tieto rezíduá nespĺňajú tieto vlastnosti, naznačuje to, že Box-Jenkinsov model nemusí byť vhodný pre dané dáta a mal by sa zvážiť vhodnejší model. V kontexte diagnostiky Box-Jenkinsových modelov sa "rezíduá" vzťahujú na rozdiely medzi skutočnými hodnotami časového radu a hodnotami predpovedanými pomocou Box-Jenkinsovho modelu [23].

3.3 Lineárne dynamické modely

Lineárne modely sú typom modelov časových radov, ktoré predpokladajú lineárny vzťah medzi minulými a súčasnými hodnotami časového radu. V tomto modeli sa predpokladá, že aktuálna hodnota časového radu závisí na lineárnej kombinácii minulých hodnôt a chybových podmienok. Predpokladá sa, že chybové podmienky majú normálne rozdelenie so strednou hodnotou nula a konštantnou varianciou. Parametre modelu sú odhadnuté pomocou štatistických metód, akou je napríklad metóda maximálnej pravdepodobnosti. Lineárne kauzálne modely sa často používajú v analýze a prognózovaní časových radov a spravidla sa jedná o príčinné (kauzálne) modely. Rozdiel od modelu Box-Jenkins spočíva v tom, že okrem popisovaného časového radu a bieleho šumu, z týchto modelov vystupujú ešte ďalšie časové rady, a tými sú príčinné faktory tiež nazývané exogénne premenné. Tieto premenné predstavujú vplyv vonkajších faktorov na modelovanú časovú radu, ktoré môžu ovplyvniť jej vývoj a správanie sa v čase. Príkladom exogénnej premennej môže byť počet dní v mesiaci alebo počasie, ak sa modeluje predaj zmrzliny. Tieto faktory sú zvyčajne preddefinované a dajú sa získať z vonkajších zdrojov, ako sú napríklad kalendáre alebo meteorologické údaje. Ich použitie môže pomôcť zlepšiť presnosť modelovania a predikcie časových radov [9, 16].

3.4 Spektrálna analýza časových radov

Predchádzajúce tri prístupy je možné zhrnúť pod označenie analýza časových radov v časovej doméne. Táto metóda pracuje s faktom, že mnohé časové rady vykazujú periodické správanie, ktoré môže byť pomerne komplexné a zložité. Spektrálna analýza je matematická technika používaná v analýze na rozloženie časových radov na základné frekvenčné zložky a umožňuje urobiť si obraz o intenzite zastúpenia jednotlivých frekvencií (tzv. spektrum radu), čo môže poskytnúť prehľad o dôležitých vzoroch a pomôcť pri predpovedaní budúcich hodnôt. Na rozdiel od predchádzajúcich troch modelov má odlišný prístup spočívajúci v tom, že sa skúmaný časový rad považuje za (nekonečnú) zmes sínusových a kosínusových kriviek s rôznymi amplitúdami a frekvenciami. Spektrálna analýza je tiež vhodná pri porovnávaní niekoľkých radov zároveň, ale tiež dovoľuje porovnať rady v rámci jednotlivých frekvencií. Jednou z najpoužívanejších metód spektrálnej analýzy v časových radoch je Fourierova transformácia, ktorá transformuje časový rad z časovej do frekvenčnej oblasti. Spektrálna analýza sa vo veľkom zastúpení používa v oblastiach, ako je spracovanie signálov, fyzika a ekonómia, a možno ju použiť na rôzne typy údajov časových radov vrátane

stacionárnych a nestacionárnych radov. Technické detaily spektrálnej analýzy však výrazne presahujú rámec tejto práce [3, 9, 14, 16].

3.5 Predpovede v časových radoch

Jednou z najdôležitejších úloh a najčastejším cieľom analýzy časových radov je predpovedanie budúcich hodnôt a analyzovanie chovania radu. Tieto predpovede sa ďalej využívajú v mnohých analýzach a úvahách o vývojoch určitých veličín.

Samozrejmosťou je, že každý takýto odhad je zaťažený chybou, a že sa jedná iba o pravdepodobný odhad a odhadovaná hodnota nemusí ani zďaleka zodpovedať skutočnosti, aj keď sa jej pri budovaní modelov snažíme čo najviac priblížiť. Je teda potrebné si uvedomiť, že hodnoty získané z odhadu sú z veľkej časti nepresné a na tieto hodnoty musíme nahliadať ako na hodnoty približné/orientačné [4].

Odhady a predikcie je možné vykonávať dvomi spôsobmi:

- Bodová predpoveď:

  Postup tejto predpovede spočíva v tom, že na základe určitých pravidiel vypočítame nasledujúcu hodnotu, o ktorej na základe modelu predpokladáme, že sa najviac približuje skutočnosti. Táto predpoveď, ako sa dá usúdiť už z názvu, vracia jedno konkrétne číslo. Pri tomto type predpovede je dôležité výslednú hodnotu brať s rezervou, pretože je vždy zaťažená chybou [4].

- Intervalová predpoveď:

  Intervalová predpoveď na rozdiel od bodovej predpovede určuje spoľahlivosť v rozmedzí intervalu pomocou matematickej štatistiky. Intervalový odhad vznikne skonštruovaním intervalu, ktorý zahŕňa hodnotu základného súboru dát a udáva hornú a dolnú hranicu, medzi ktorými bude ležať budúca hodnota sledovaného časového radu s určitou spoľahlivosťou. Spoľahlivosť je definovaná ako $(1 - \alpha)$, kde α je hladina významnosti pohybujúca sa medzi číslom 0 a 1 [4].

Ďalej sa hodnoty pre predpovedanie delia na:

- Kvalitatívne predpovedné metódy

  Tieto metódy sú využívané v prípade, že nie sú k dispozícii predchádzajúce historické dáta. Sú teda najčastejšie založené na názore odborníka [4].

- Kvantitatívne predpovedné metódy

  Predpovede týchto metód sú založené na základe štatistickej analýzy predchádzajúcich údajov. Kvantitatívne predpovedné metódy totiž vykonávajú autoprojekciu, predĺženie (inými slovami extrapoláciu) prítomných a minulých hodnôt časového radu do budúcnosti. Z toho vyplýva, že tieto metódy počítajú s tým, že daný rad v budúcnosti (pre ktorú robíme predpoveď) nezmení svoj aktuálny charakter [4].

3.6 Základné algoritmy pre tvorbu modelu

Táto kapitola priblíži niektoré základné algoritmy používajúce sa pre tvorbu modelov predikcie, ktoré sú stavebnými kameňmi tejto disciplíny a zohrávajú kľúčovú úlohu pri pochopení jej konceptov. Sú navrhnuté tak, aby zachytili základné vzorce a vzťahy v časových radoch a umožňujú vytvárať predikcie hodnôt do budúcnosti.

3.6.1 AR modely

Auto-regresný model časového radu je založený na poznatku, že každá hodnota časového radu je v relácii s predchádzajúcimi hodnotami tohto časového radu a robí túto štatistickú techniku populárnou pre analýzu a predpovedanie ekonomiky alebo procesov v iných odvetviach, ktoré sa menia v čase. Druhou dôležitou vlastnosťou týchto modelov, vďaka ktorej je táto metóda pomerne rozšírená je, že sú pozoruhodne flexibilné pri manipulácii so širokým spektrom rôznych vzorov časových radov a dajú sa využiť na veľkú škálu dát.

Aktuálna hodnota časového radu sa predpovedá pomocou lineárnej kombinácie minulých hodnôt, ak sa jedná o auto-regresný model, v prípade viacnásobného regresného modelu sa hodnoty predpovedajú na základe lineárnej hodnoty prediktorov. Pojem auto-regresia značí, že ide o regresiu premennej voči sebe samej.

Pre priblíženie si môžeme uviesť, že auto-regresný proces AR(1) je proces, v ktorom je aktuálna hodnota založená na bezprostredne predchádzajúcej hodnote, zatiaľ čo proces AR(2) je proces, v ktorom je aktuálna hodnota založená na dvoch predchádzajúcich hodnotách.

Auto-regresný model AR(p), ktorý má rad p, je zapísaný nasledujúcou rovnicou [4]:

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t,$$

kde

$\varphi_1, \varphi_2, \ldots, \varphi_p$ sú koeficienty autoregresného procesu

ε_t je biely šum (proces s nulovou strednou hodnotou, konštantným rozptylom a nulovou ACF a PACF)

y_t je nová hodnota radu vypočítaná na základe predchádzajúcich hodnôt

Zmena parametrov $\varphi_1, \ldots, \varphi_p$ vyústi v zmenu vzorov časového radu. Ale zmena parametru ε_t zmení iba mierku časového radu, nie je vzor.

Tento proces sa dá zapísať aj pomocou symboliky operátoru spätného posunutia B ako [4]:

$$\varphi(B) Y_t = \varepsilon_t,$$

kde

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p,$$

je tzv. auto-regresný operátor [16].

Proces AR(p) je stacionárny, ak všetky jeho koeficienty polynómu $\varphi(B)$ ležia vnútri jednotkového kruhu (v komplexnej rovine).

Auto-korelačná funkcia AR(p) procesu má tvar [6]:

$$\rho_k = \alpha_1 G_1^{-k} + \cdots + \alpha_p G_p^{-k} \qquad k \geq 0$$

kde

α sú konštanty

G sú korene polynómu $\varphi(B)$, ktoré sú navzájom rôzne a ležia vnútri jednotkového kruhu

Auto-korelačná funkcia auto-regresného procesu rádu p je teda lineárnou kombináciou klesajúcich geometrických postupností a sínusoíd rôznych frekvencií s geometricky klesajúcimi amplitúdami.

Pre výpočet parametrov auto-regresného procesu rádu p pomocou hodnôt jeho auto-korelačnej funkcie sa používa tzv. Yuelova-Walkerova sústava lineárnych rovníc [2, 3, 4, 5, 6, 12, 16].

3.6.2 MA modely

Na začiatku je nutné upozorniť, že sa jedná o proces kĺzavých súčtov radu q a nie o metódu kĺzavých priemerov používanou pre elimináciu trendu. Tento model namiesto regresie minulých hodnôt časového radu, využíva odchýlky hodnôt minulých predpovedí (tiež nazývaných rezíduá) v modeli podobnom regresii. Model konkrétne predpokladá, že aktuálna hodnota premennej závisí od priemeru časového radu plus lineárnej kombinácie predchádzajúcich rezíduí.

Podobne ako pri AR modeloch, rád modelu MA sa vzťahuje na počet predchádzajúcich rezíduí použitých v modeli. A teda $MA(1)$ model používa len najnovšiu chybovú odchýlku, zatiaľ čo model $MA(2)$ používa dva najnovšie chybové výrazy.

Proces kĺzavých súčtov radu q značený ako $MA(q)$ (z anglického Moving Average) má tvar [4]:

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

kde ε_t je biely šum. Je samozrejmosťou, že sa nesledujú hodnoty ε_t, preto sa tento proces nedá považovať za regresiu v bežnom slova zmysle.

Rovnako sa tento model dá zapísať pomocou symboliky operátoru spätného posunutia B ako [4]:

$$y_t = \theta(B)\varepsilon_t,$$

kde

$$\theta(B) = 1 + \sum_{j=1}^{q} \theta_j B^j$$

je tzv. operátor kĺzavých súčtov. Pritom ε_t je biely šum a θ_j sú parametre modelu. Z definície vieme, že proces $MA(q)$ je stacionárny pre ľubovoľnú voľbu parametrov.

Zaujímavou vlastnosťou je, že každý stacionárny auto-regresný proces rádu $p(AR(p))$ je možné modelovať ako proces kĺzavých súčtov rádu nekonečno $(MA(\infty))$, napríklad pomocou opakovanej substitúcie. Naopak to platí len v prípade, že zavedieme určité obmedzenia pre parametre procesu kĺzavých súčtov. Takýto model MA následne nazývame invertovateľný. To znamená, že hocijaký invertovateľný proces $MA(q)$ sme schopný zapísať ako proces $AR(\infty)$.

Modely MA sa často používajú zároveň s AR modelmi, kde sa spájajú a vytvárajú tzv. modely ARMA [2, 3, 4, 5, 6, 12, 16].

3.6.3 ARMA model

Pre zopakovanie, kombináciou procesov $AR(p)$ a $MA(q)$ vzniká zmiešaný proces $ARMA(p, q)$. Rád modelu ARMA tvoria dva parametre: rád auto-regresívnej zložky (p) a rád zložky

kĺzavých súčtov (q). Kde zložka AR modelu zohľadňuje závislosť aktuálnej hodnoty od jej predchádzajúcich hodnôt až po oneskorenie p, zatiaľ čo zložka MA zohľadňuje závislosť od predchádzajúcich rezíduí po oneskorenie q.

Proces $ARMA(p, q)$ sa tiež nazýva zmiešaný a je definovaný ako [16]:

$$y_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon t - 1 + \cdots + \theta_q \varepsilon t - q,$$

alebo pomocou symboliky operátoru spätného posunutia B ako [16]:

$$\varphi(B) y_t = \theta(B) \varepsilon_t,$$

kde $\varphi(B)$ a $\theta(B)$ boli definované v predchádzajúcich dvoch kapitolách [16]. Podmienka stacionarity tohto procesu je totožná s podmienkou stacionarity procesu $AR(p)$ a podmienka inverovateľnosti je totožná s podmienkou invertovateľnosti procesu $MA(q)$. Stredná hodnota procesu $ARMA$ je nulová a jeho auto-korelačná funkcia zodpovedá podobnej sústave diferenčných rovníc ako v prípade AR procesu [16].

3.6.4 ARIMA modely

Tieto modely slúžia pre popis procesov, pri ktorých nielenže dochádza k zmene úrovne, ale tieto zmeny môžu mať náhodný nesystematický charakter, ako je to bežné pre veľké množstvo časových radov. Tieto modely zvládajú modelovať stochasticky okrem náhodných fluktuácií aj trendovú zložku, čo umožňuje modelovať aj časové rady ktoré nie sú stacionárne, avšak je nutné, aby sa tieto rady dali previesť na stacionárne. Tento prevod sa uskutočňuje diferencovaním pôvodného časového radu [4, 16].
Integrovaný zmiešaný model $ARIMA(p, d, q)$ sa definuje ako [4]:

$$\varphi(B) w_t = \theta(B) \varepsilon_t,$$

kde

$$w_t = \Delta^d y_t$$

W_t reprezentuje časovú radu skonštruovanú diferenciáciou pôvodného radu Y_t, d je rád diferencovania a delta diferenčný operátor definovaný ako [4]:

$$\Delta^d Y_t = (1 - B)^d Y_t$$

Tento model sa teda tiež dá napísať v tvare [4]:

$$\varphi(B)(1 - B)^d Y_t = \theta(B) \varepsilon_t.$$

Konštrukcia zmiešaného integrovaného modelu sa realizuje v dvoch krokoch [16]:

1. Prvým krokom je prevod nestacionárneho časového radu W_t (pomocou vhodnej transformácie) na stacionárny rad Y_t. Pri tomto kroku je potrebné si uvedomiť, že diferencovaním sa pôvodný časový rad skracuje.

2. Na transformovanú stacionárnu radu sa následne v druhom kroku aplikuje zmiešaný model $ARMA(p, q)$.

3.6.5 SARIMA modely

Sezónne zmiešané integrované modely SARIMA slúžia k popisu časových radov v ktorých trend a sezónna zložka majú stochastický charakter [16].To znamená, že na rozdiel od zmiešaného integrovaného procesu ARIMA, ktorý umožňuje modelovať nestacionárne časové rady s trendom, model SARIMA umožňuje modelovať časové rady, ktoré okrem trendu obsahujú aj sezónnu zložku, čo je veľkým prínosom napríklad pre dáta z ekonomickej sféry, ktoré môžu zvyčajne v rámci určitej sezóny (napr. jeden rok) opakovať svoj vzor. Podobne ako v prípade sezónneho integrovaného procesu ARIMA sa predpokladá vzájomná závislosť medzi jednotlivými veličinami časového radu a zároveň sa očakáva aj závislosť medzi sebe odpovedajúcimi veličinami v jednotlivých sezónach, nakoľko tento proces naviac obsahuje sezónne kolísanie. To sa dá priblížiť ako závislosť medzi veličinami ... $y_{t-2s}, y_{t-1s}, y_{ts}, y_{t+1s}, y_{t+2s}, \ldots$, kde s je dĺžka sezónnej periódy [2].

Ak je teda predpokladom, že tento proces obsahuje oba spomenuté typy závislostí. Závislosť v rámci období je zachytená modelom ARIMA [2]:

$$\phi_P(B)(1-B)^d y_t = \theta_q(B) b_t$$

Proces $\{bt\}$ obsahuje len sezónnu závislosť a môže byť popísaný modelom [2]:

$$\Phi_P(B^s)(1-B^s)^D b^t = \Theta_Q(B^s) a_t$$

kde

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \ldots - \Phi_P B^{Ps}$$
$$\Theta_Q(B^s) = 1 - \Theta_1 B^s - \ldots - \Theta_Q B^{Qs}$$

Prostredníctvom člena $(1-B^s)$ sa konštruujú sezónne diferencie. Ak sa spoja, prvé dva popísane procesy v tejto kapitole vzniká proces [2]:

$$\Theta_Q(B^s)\phi_P(B)(1-B)^d y_t = \theta_q(B)\Theta_Q(B^s) a_t,$$

ktorý sa označuje ako SARIMA(p,d,q)(P,D,Q)s, kde z predchádzajúcich kapitol p je rád procesu AR, d je rád diferencie, q rád procesu MA, P rád sezónneho procesu AR, D rád sezónnej diferencie, Q rád sezónneho procesu MA a s je dĺžka sezónnej periódy. Podmienky stacionarity a invertovateľnosti nesezónnej zložky, vychádzajú z podmienok modelu ARIMA, a podmienky stacionarity a invertovateľnosti sezónnej zložky sú rovnaké ako pri zložke nesezónnej [2].

3.7 Identifikácia modelu

Najväčšou výzvou pri výstavbe modelu pre konkrétne časové rady, je identifikácia týchto modelov. Hlavná úloha spočíva v rozhodnutí, ktorý typ modelu bude najvhodnejší a teda ktorý typ pre konkrétny časový rad vybrať. Ide pomerne o zložitú činnosť, ktorá vo veľkej miere závisí na skúsenostiach a cite analytika. Je potrebné si uvedomiť, že identifikácia modelu je len prvou fázou konštrukcie modelu, pretože identifikovaný model je treba následne overiť a upraviť, aby sa čo najviac priblížil skutočnosti. Nižšie sú popísané jednotlivé kroky identifikácie modelu [2].

Uvažujme model ARIMA(p,d,q) ktorý bol popísaný v kapitole 3.6.4.

1. Ako prvým krokom pri identifikácii modelu je vhodné preskúmať graf časového radu, z ktorého je častokrát na prvý pohľad možné rozpoznať prítomnosť trendu (príp. sezónnej zložky). Treba však podotknúť, že v tejto fáze ide predovšetkým o subjektívne

vyhodnotenie situácie, nakoľko nie vždy musia byť jednotlivé zložky jednoznačné. Napriek tom, už na základe tohto vyhodnotenia je možné stacionarizovať časovú radu alebo stabilizovať ju z hľadiska rozptylu pomocou logaritmickej funkcie. Avšak treba podotknúť, že takýto typ transformácie je vhodné vykonať pred vlastným diferencovaním časového radu, ktorý môže vrátiť aj záporné hodnoty [2].

2. Druhým krokom je výpočet odhadov auto-korelačnej (ACF) a parciálne autokorelačnej funkcie (PACF) pôvodného časového radu. Na základe ktorých, je možné zistiť či je potrebné časový rad stacionarizovať (v prípade, že sú hodnoty výberovej ACF A PACF v prvom oneskorení veľmi blízko jednej a ostatné hodnoty výberovej ACF pomaly klesajú) [2]. Tu sa nám rovnako ponúka využiť, jeden zo štatistických testov (ADF, KPSS) na pretestovanie stacionárnosti časového radu z kapitoly 2.5.3.

3. Po prevedení nestacionárneho časového radu na stacionárny sa používajú výberové ACF a PACF pre identifikáciu modelov AR a MA, teda určenie hodnôt p a q. Táto identifikácia je založená na princípe podobností výberových ACF a PACF s teoretickými ACF A PACF [2].

Ak však časový rad obsahuje aj sezónnu zložku, ktorá by mala byť zvyčajne rozpoznateľná už z grafu, je potrebné identifikovať model SARIMA. Princíp je podobný ako v predchádzajúcom prípade.

1. Obdobne ako pri modeli ARIMA je nutné najprv preskúmať graf časového radu a časový rad stacionarizozvať. Ak je to nutné, rad sa najprv linearizuje za pomoci logaritmickej transformácie. Následne sa rad diferencuje, avšak je potreba dať si pozor na fakt, že pri tomto druhu modelu, sezónna diferencia zahŕňa aj prostú diferenciu. Stacionarizácia pomocou sezónnej diferencie indikuje tvar výberovej ACF a PACF, kde sú tieto funkcie charakteristické vysokými hodnotami v nesezónnych a sezónnych frekvenciách [2].

2. V druhom kroku sa opätovne vypočítajú výberové ACF a PACF pre už stacionarizovanú časovú radu, vďaka ktorým sa následne určí typ sezónneho modelu ktorého funkcie ACF a PACF majú v sezónnych frekvenciách štatisticky významné odlišné hodnoty od nuly. Tieto hodnoty však nie sú tak vysoké, aby bolo možné časovú radu považovať za nestacionárnu [2].

Kapitola 4

Charakteristika dát

Táto kapitola sa venuje popisu a prvotnej analýze dátovej sady na ktorej stojí tvorba tejto práce. Sada bola zozbieraná z rôznych priemyselných zdrojov a rôznych typov komunikácie, ako bude popísané v kapitole neskôr. Dáta boli vopred predspracované a agregované na základe ich charakteristiky do odpovedajúcej skupiny a príslušného časového okna, kde jedno okno predstavuje päť minútový interval, a tieto dáta sú uložené vo formáte CSV (teda "hodnoty oddelené bodkočiarkou"). Každá dátová sada sa skladá zo šiestich stĺpcov (časových radov) a čísla okna, čo popisuje obrázok 4.1.

	fm_sum	fm_under	fm_above	tm_sum	tm_under	tm_above
0	63	56	7	24	9	15
1	67	62	5	23	4	19
2	59	47	12	24	10	14
3	63	52	11	25	8	17
4	54	45	9	24	7	17

Obr. 4.1: Zobrazenie hlavičky a formátu dátovej sady

Stĺpce **fm_sum** a **tm_sum** predstavujú celkový počet paketov, kde značky fm a tm predstavujú smer komunikácie :

fm: from-master (komunikácia odchádzajúca z nadradeného zariadenia)

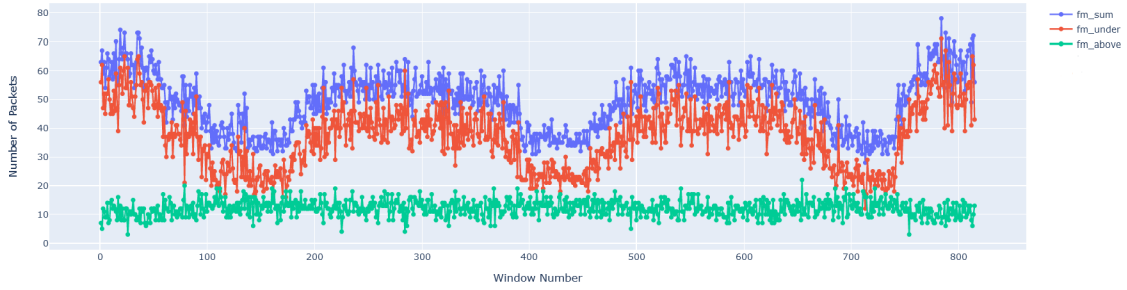
tm: to-master (komunikácia prichádzajúca k nadradenému zariadeniu)

Toto značenie je založené na fakte, že všetka komunikácia v týchto sieťach, bola z hľadiska modelovania delená do dvoch smerov a to od master stanice a k master stanici (zvyčajne sa komunikácia v sieťovej premávke rovnako delí na prichádzajúce a odchádzajúce dáta). V prípade, že sa jedná o komunikáciu medzi jednou nadradenou stanicou (master) a niekoľkými podriadenými stanicami (slave), tak sa v oknách fm nachádzajú všetky pakety akumulované do jedného čísla, ktoré odchádzajú od nadradenej stanice komukoľvek a v tm sa nachádzajú všetky pakety, ktoré k nadradenej stanici prichádzajú od kohokoľvek.

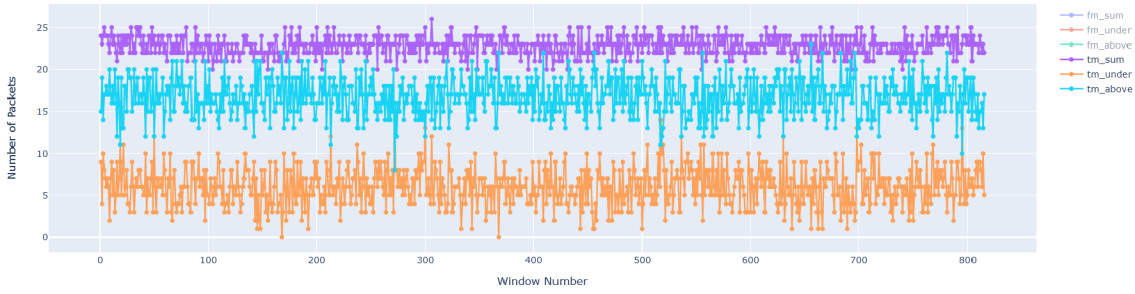
Hodnoty above a under, sú založené na dodatočnom delení paketov prenesených v jednotlivých smeroch, na základe ich medzipaketového oneskorenia. Kde sa stanovil hraničný bod, a pakety, ktoré dorazili za kratší čas, sa objavili v stĺpci *fm_under* príp. *tm_under* (záleží od smeru komunikácie), zvyšné v stĺpcoch *fm_above* a *tm_above*. V rámci tejto práce boli zvolené za hlavné ukazovatele stĺpce *fm_sum* a *tm_sum* s ktorými sa následne pracovalo.

Dáta, ktoré boli v tejto práci použité, sa skladajú zo štyroch dátových sád zachytávajúcich normálnu premávku určených pre trénovanie, k nim príslušným trom dátovým sadám určeným pre testovanie (testovacia časť pre jednu dátovú sadu chýbala, preto ju bolo potrebné vytvoriť ručne rozdelením danej dátovej sady na trénovaciu a testovaciu časť) a šiestich časových radov, ktoré majú v sebe simulovaný nejaký typ útoku.

Jedným z najdôležitejších časových radov tejto práce, je časový rad *normal-traffic-win5-3v-nb.csv*, ktorý je vyobrazený na obrázku 4.2, ktorý vykresľuje hodnoty stĺpcov fm, a obrázku 4.3, ktorý vykresľuje stĺpce tm. Jedná sa o normálnu IEC 104 komunikáciu v ktorej bolo prenesených dohromady 58930 paketov za celkový čas 2 dni 19 hodín a 55 minút, čo v prípade agregácie na 5 minútové okná, činí 815 okien. Ďalšou vlastnosťou tohto radu, ktorý sa po preskúmaní grafov odhalí na prvý pohľad, je sezónnosť radov *fm_sum* a *fm_under*, z čoho sa dá vyvodiť nesplnenie podmienky stacionarity tohoto radu. Zatiaľ čo zvyšné štyri časové rady/stĺpce sa na prvý pohľad javia bez stacionarity, trendu a variancia sa zdá konštantná v čase. Za najdôležitejšiu časovú radu, sa dá považovať kvôli faktu, že práve do tohto konkrétneho časového radu boli simulované všetky útoky a anomálie, s ktorými sa bude v ďalších častiach ďalej pracovať. To činí šesť ďalších časových radov, kde každý z nich má rovnakú časovú dĺžku ako časový rad *normal-traffic-win5-3v-nb.csv*.

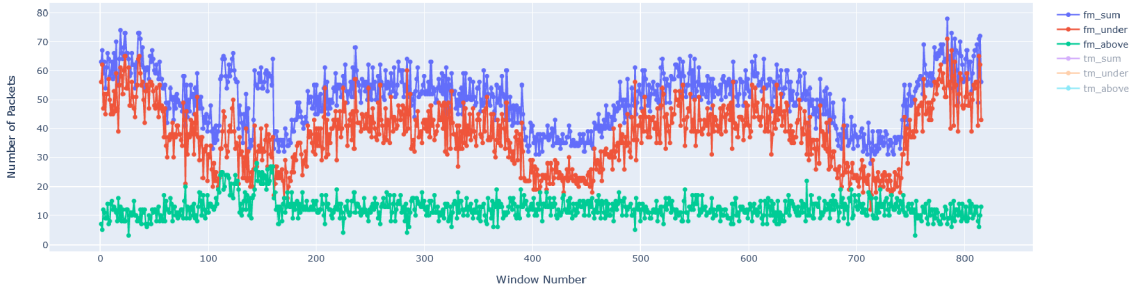


Obr. 4.2: Graf dátovej sady normal-traffic-win5-3v-nb.csv v smere od master-a

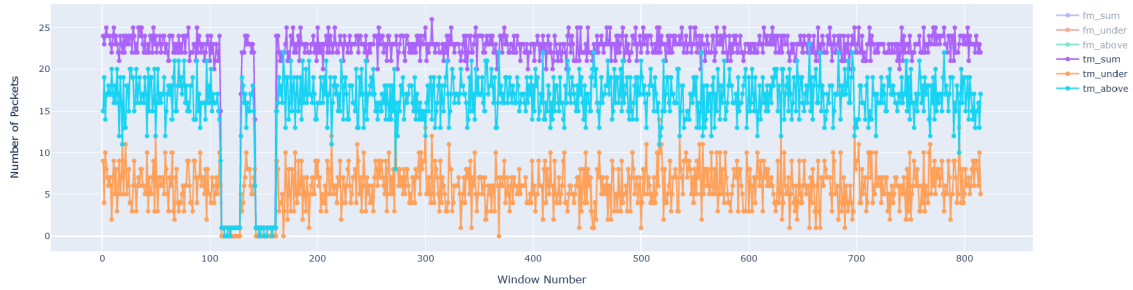


Obr. 4.3: Graf dátovej sady normal-traffic-win5-3v-nb.csv v smere k master-ovi

Prvým časovým radom zo série simulovaných útokov, je časový rad *dos-attack-win5-3v-nb.csv*. Do tohto časového radu bol simulovaný DOS útok, ktorý je možné pozorovať na obrázkoch 4.4 a 4.5. Cieľom DOS útoku je spomaliť alebo zhodiť danú službu, a to záplavou cieľa nadbytočnou premávkou alebo odoslaním dát, ktoré vyústia v zlyhanie systému. Útočník v tomto prípade odoslal do cieľovej stanice stovky legitímnych paketov IEC 104 a to v oknách 110 až 128 a 142 až 161.

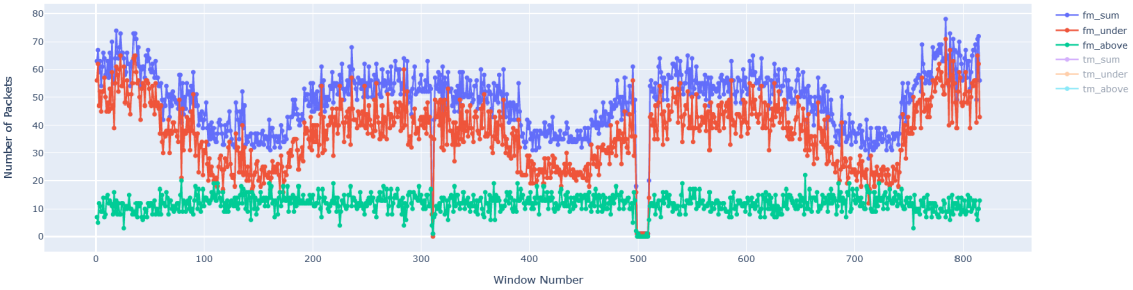


Obr. 4.4: Graf dátovej sady so simulovaným DOS útokom v smere od master-a

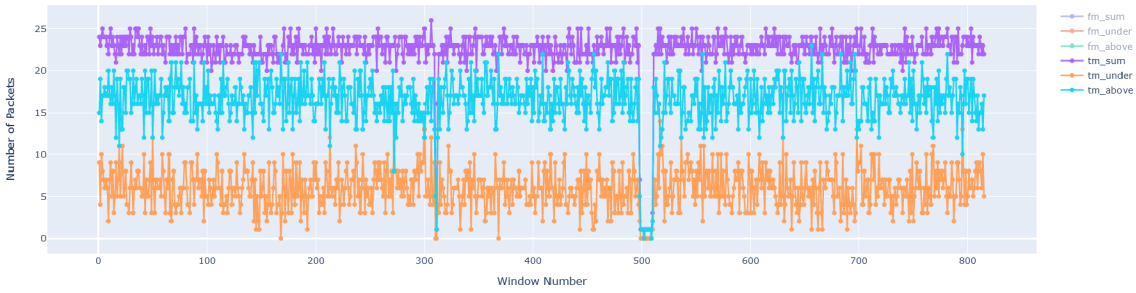


Obr. 4.5: Graf dátovej sady so simulovaným DOS útokom v smere k master-ovi

Nasledujúcim časovým radom, je rad *connection-loss-win5-3v-nb.csv*, v rámci ktorého, ako je vidieť na obrázkoch 4.6 a 4.7, došlo dvakrát k výpadku spojenia medzi nadradenou stanicou a stanicami, ktoré sa na ňu pripájajú. Prvý výpadok trval 10 minút (okná 310 až 312) a došlo k strate 146 paketov, druhý výpadok, počas ktorého došlo k strate 921 paketov, bol dlhší a trval až 1 hodinu (okná 498 až 510).

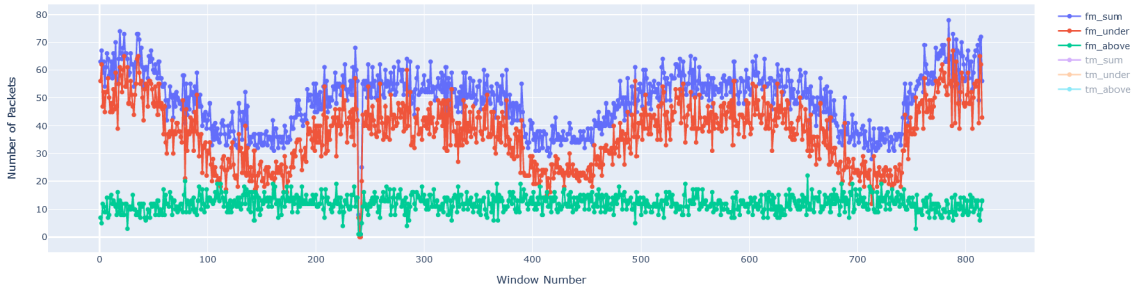


Obr. 4.6: Graf dátovej sady so simulovaným výpadkom v smere od master-a

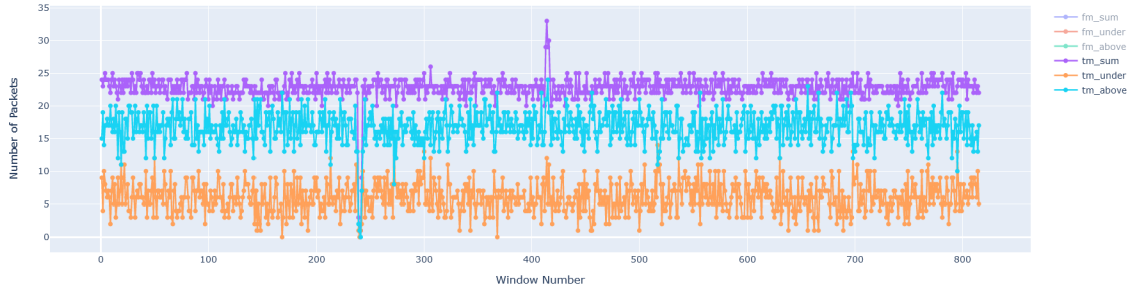


Obr. 4.7: Graf dátovej sady so simulovaným výpadkom v smere k master-ovi

Tretím časovým radom simulovaných útokov, je časový rad *scanning-attack.csv*, ktorý obsahuje typ útokov nazývajúcich sa horizontálne a vertikálne skenovanie vyobrazených na obrázku 4.8 a obrázku 4.9. Hlavným rozdielom týchto dvoch typov, je ich charakter akým sa skenovanie vykonáva. Pri horizontálnom skenovaní sa skenuje skupina IP adries (niekoľko zariadení) pre jeden konkrétny port. Naopak pri skenovaní vertikálnom dochádza k skenovaniu jednej adresy IP (jedného zariadenia) pre viacero (skupinu) portov. Útok horizontálneho skenovania sa nachádza v oknách 239 až 242 a vertikálne skenovanie sa nachádza v oknách 413 až 417.



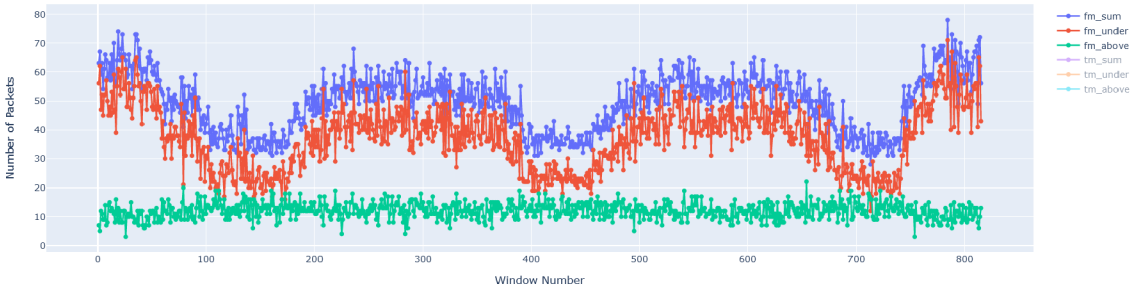
Obr. 4.8: Graf dátovej sady so simulovaným skenovaním v smere od master-a



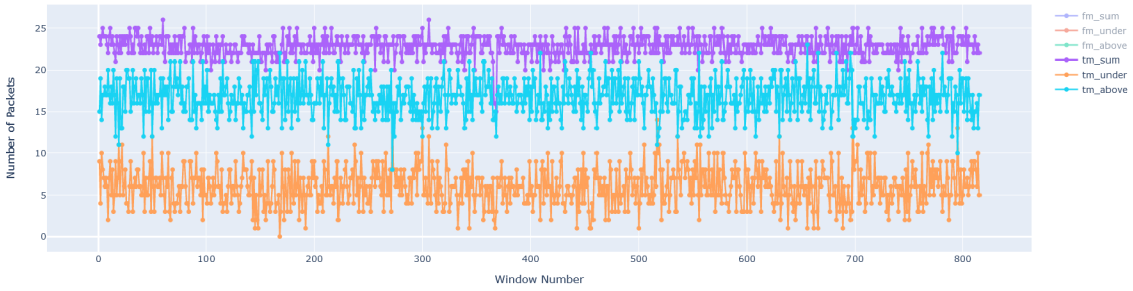
Obr. 4.9: Graf dátovej sady so simulovaným skenovaním v smere k master-ovi

Ďalším typom útoku, ktorý dátová sada zahŕňa, je tzv. “injection attack” ktorý sa ako súčasť dotazu (alebo príkazu), snaží vložiť nedôveryhodné vstupy alebo neautorizovanú časť kódu, ktorá môže spôsobiť rôzne škody príp. pád systému. Táto dátová sada sa nachádza v súbore injection-attack-win5-3v-nb.csv a bližšie ju možno preskúmať na obrázkoch 4.10 a 4.11. Kde v časových oknách 59 až 60 útočník posiela nezvyčajné požiadavky a v oknách

365 až 368 útočník odosiela súbor kompromitovanému zariadeniu, vďaka ktorému pristúpi k objektu, ktorý by nemal byť normálne prístupný.

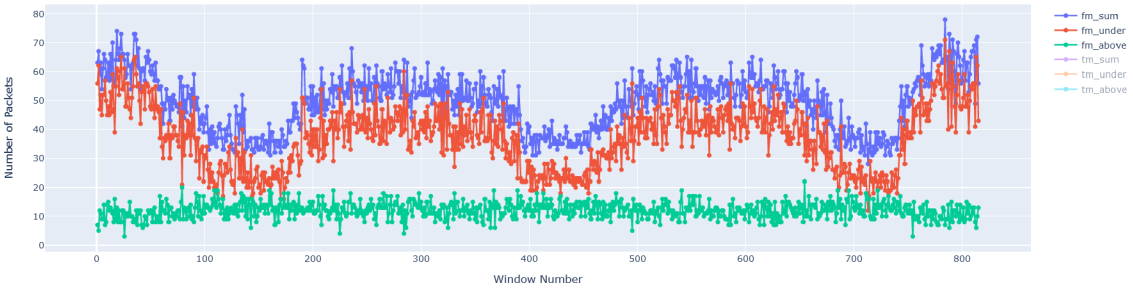


Obr. 4.10: Graf dátovej sady so simulovanou injekciou v smere od master-a

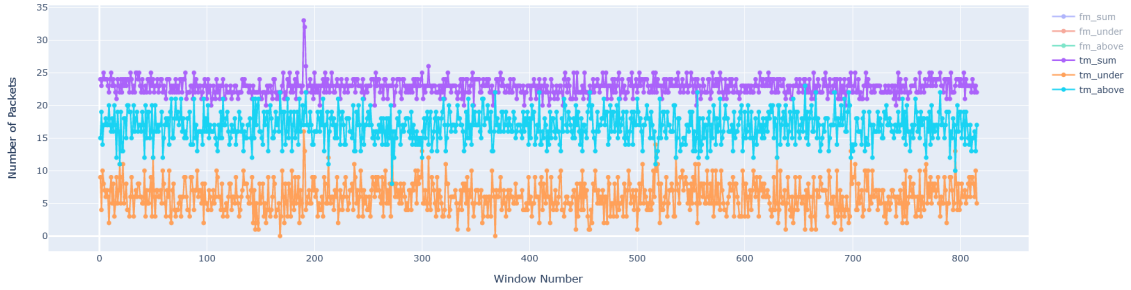


Obr. 4.11: Graf dátovej sady so simulovanou injekciou v smere k master-ovi

Predposledným časovým radom, zo série simulovaných útokov je *switching-attack-win5-3v-nb.csv* viditeľný na obrázkoch 4.12 a 4.13. V tomto časovom rade je na rozdiel od predchádzajúcich, simulovaný len jeden útok, a jedná sa o tzv. switching útok pri ktorom útočník vypínal a zapínal zariadenie, čím sa vygenerovalo sedemdesiatdva nových paketov. Tento útok, vidieť v oknách 190 až 192.

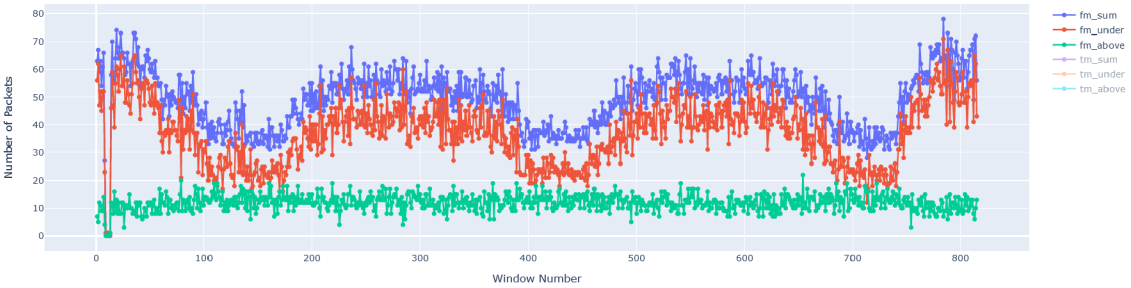


Obr. 4.12: Graf dátovej sady so switching útokom v smere od master-a

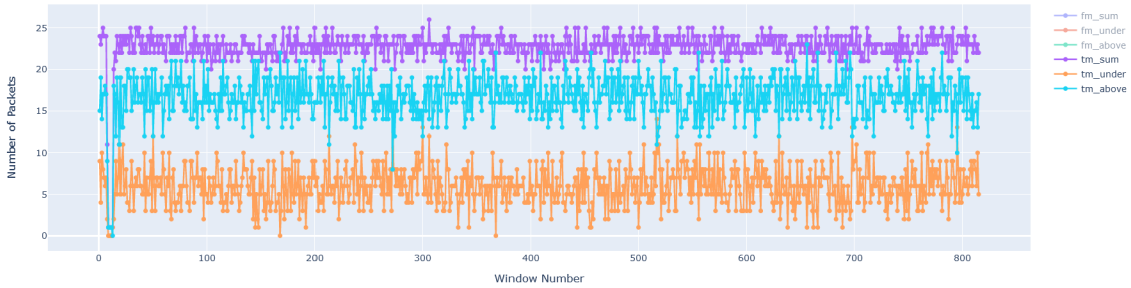


Obr. 4.13: Graf dátovej sady so switching útokom v smere k master-ovi

Poslednou anomáliou, zo série šiestich simulovaných útokov je tzv. Rogue device nachádzajúci sa v súbore raw-device-win5-3v-nb.csv. Jedná sa o neoprávnené pripojenie zariadení k sieti, ktoré predstavujú bezpečnostnú hrozbu pre organizáciu. Pripojenie tohto zariadenia je možné pozorovať v oknách 8 až 13 na obrázkoch 4.14 a 4.15.



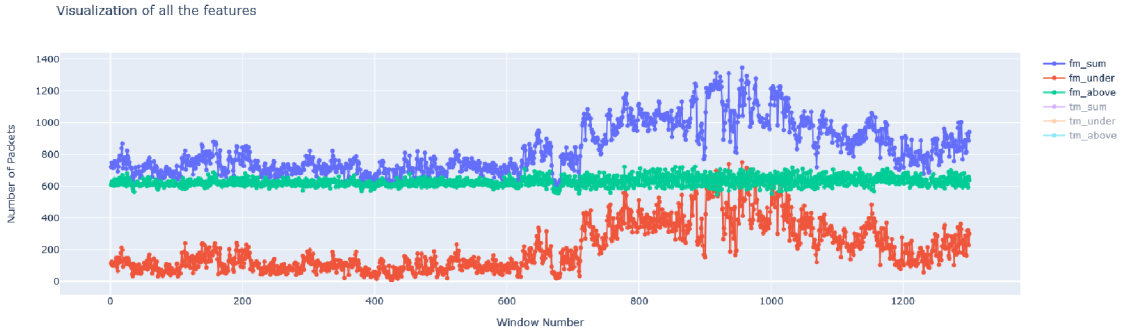
Obr. 4.14: Graf dátovej sady s rogue device útokom v smere od master-a



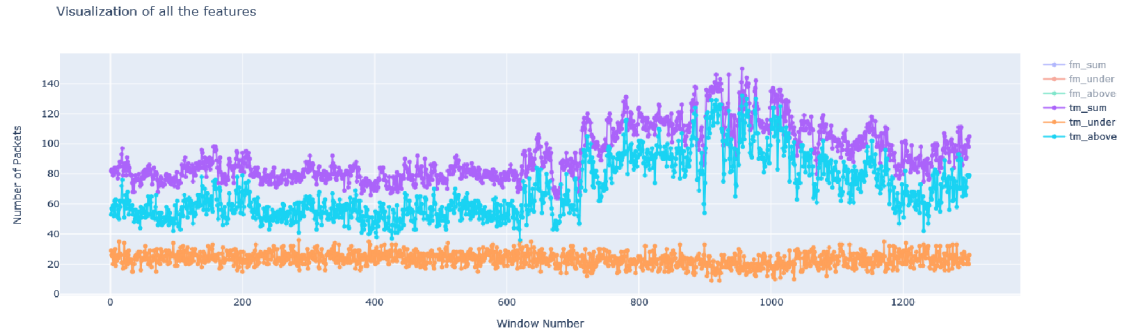
Obr. 4.15: Graf dátovej sady s rogue device útokom v smere k master-ovi

Na záver budú popísané zostávajúce tri časové rady, ktoré zachytávajú bežnú internetovú premávku bez simulovaných útokov a anomálií. Tieto tri časové rady boli vopred rozdelené na trénovaciu a testovaciu časť, preto v tejto kapitole budú popísané a zobrazené len častí určené pre trénovanie.

Na obrázkoch 4.16 a 4.17 je vyobrazený časový rad *rtu11-train.csv* skladajúci sa z 1300 okien, čo činí 4 dni 12 hodín a 20 minút záznamov (testovacia sada obsahuje ďalších 650 okien). Pri preskúmaní grafov vidíme že časové rady fm_sum a tm_sum majú podobný charakter. Obidve rady na prvý pohľad nevykazujú prítomnosť sezónnej zložky avšak je možné pozorovať istý v čase narastajúci trend a nekonštantnú varianciu (rozptyl). Z vizuálnej analýzy sa teda vyvodí záver, že tieto dva časové rady nie sú stacionárneho charakteru.



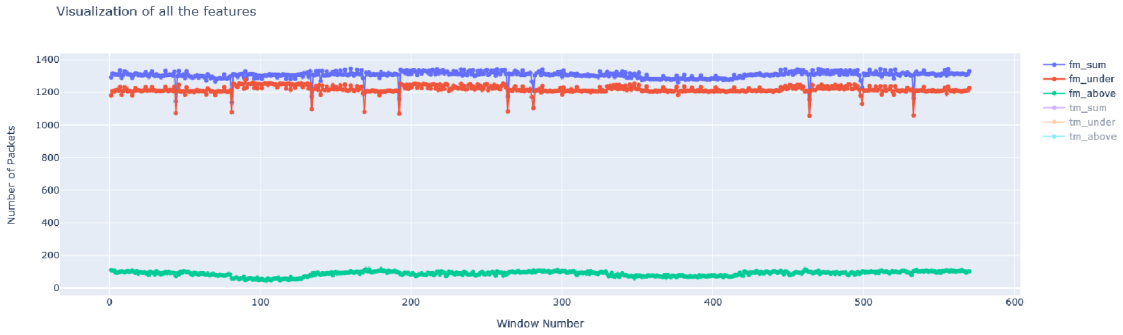
Obr. 4.16: Graf dátovej sady rtu11-train.csv v smere od master-a



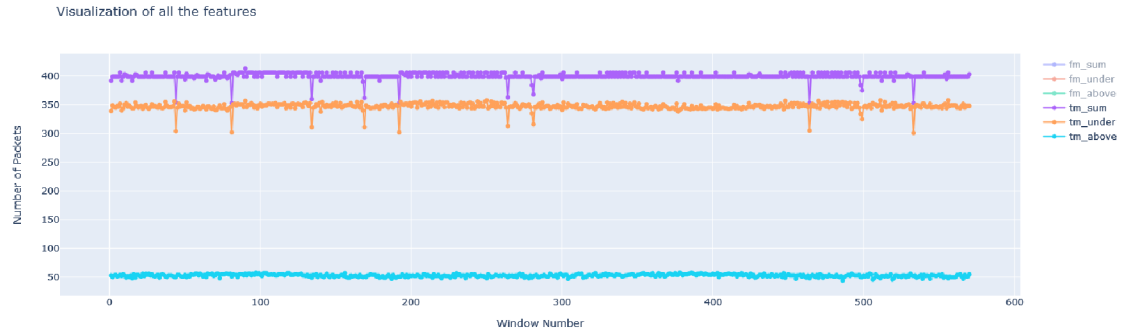
Obr. 4.17: Graf dátovej sady rtu11-train.csv v smere k master-ovi

Na ďalších dvoch obrázkoch, teda na obrázku 4.18 a obrázku 4.19 je možné vidieť časový rad *13122018-104Mega-train.csv*, ktorý sa na prvý pohľad odlišuje istými nepravidelnými prepadmi v komunikácii, čo by sa niekomu mohlo javiť ako výpadky, prípadne chyba. Tieto prepady sú však spôsobené prepájaním master stanice na druhú podriadenú stanicu s ktorou bude komunikovať. Táto zmena trvá istý čas, a preto je v daných časových oknách zachytených menej prichádzajúcich a odchádzajúcich paketov. Tieto prepady sú nepravidelné na základe faktu, že čas ktorý komunikuje master s podriadenou stanicou, je pre každú stanicu odlišný. Z analýzy grafov vidíme, že fm_sum a tm_sum majú opäť podobný charakter. Rozdielom je, že pri týchto dvoch časových radoch sa odhaduje stacionárny charakter, nakoľko sa z vizuálnej analýzy zdá, že časové rady majú konštantný priemer a rozptyl (až na

zanedbateľné výkyvy pri prepájaní) a nie je možné pozorovať žiadnu sezónnu zložku. Tento časový rad je dlhý 570 okien a testovacia sada obsahuje ďalších 285 okien.



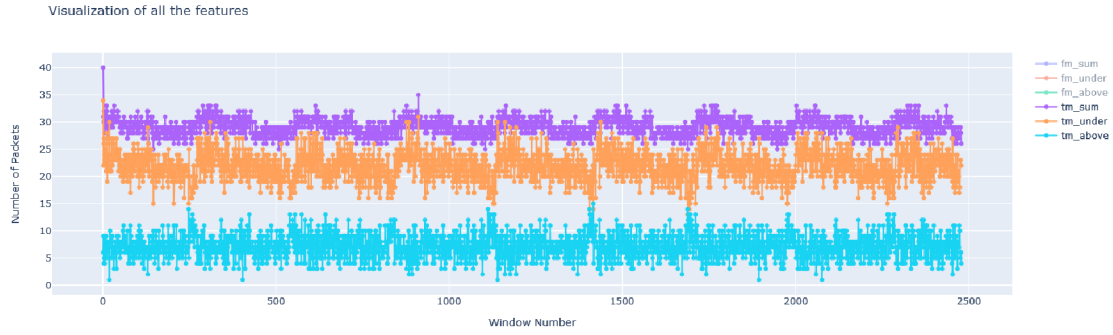
Obr. 4.18: Graf dátovej sady 13122018-104Mega-train.csv v smere od master-a



Obr. 4.19: Graf dátovej sady 13122018-104Mega-train.csv v smere k master-ovi

Kapitolu zakončí posledný časový rad a to rad *10days-train.csv*, ktorý má na prvý pohľad veľmi pekne a jednoducho pozorovateľnú sezónnosť, ktorú je možné všimnúť si na obrázkoch 4.20 a 4.21. Na základe tohto faktu sa dá rovno usúdiť, že sa bude jednať o nestacionárne časové rady a teda bude potrebné tieto rady upraviť. Svojou dĺžkou sa jedná o najdlhšiu dátovú sadu, ktorá sa skladá z 2478 okien určených pre trénovanie a ďalších 1238 okien určených pre testovanie.

Obr. 4.20: Graf dátovej sady 10days-train.csv v smere od master-a



Obr. 4.21: Graf dátovej sady 10days-train.csv v smere k master-ovi

Kapitola 5

Návrh a implementácia

Dátová analýza a analýza časových radov je komplexný proces, ktorý zahŕňa celý rad techník od vizualizácie údajov a štatistickej analýzy až po strojové učenie a prediktívne modelovanie. Táto kapitola bližšie popíše návrh a implementáciu tohto procesu pre už spomenuté časové rady pomocou skriptovacieho jazyka Python[1], Jupyter Notebook-u[2] a rôznych knižníc určených práve pre analýzu časových radov.

Skriptovací jazyk Python sa v posledných rokoch stáva čoraz populárnejším, či už v oblasti programovania alebo dátovej analýzy, vďaka jeho jednoduchému použitiu, výkonným a rozsiahlym knižniciam a aktívnej komunite. Táto práca využíva prevažne prostredie Jupyter Notebook-u, ktoré sa dá popísať ako interaktívne webové vývojové prostredie. Prostredie, ktoré si získalo obľubu medzi dátovými analytikmi a dátovými výskumníkmi pre svoju schopnosť kombinovať kód, vizualizácie a naratívny text do jedného dokumentu. Konkrétne sa práca a výsledky tejto práce nachádzajú v piatich separátnych notebookoch. Štyri notebooky sa venujú jednotlivým časovým radom s normálnou premávkou, ich analýze a predikcii. Sú štruktúrované tak, aby postup práce nasledoval princíp navrhnutý a popísaný v kapitole 5.1. Posledný notebook sa venuje dátovým sadám s útokmi, predikcii a detekcii týchto anomálii v časovom rade. Prístup v tomto súbore sa istým spôsobom odlišuje od predchádzajúcich štyroch, kde na základe poznatkov získaných z analýzy, sa už len vytvoria modely pre jednotlivé typt útokov a prebehne proces detekcie anomálií, ktorému sa venuje podkapitola 5.2.6.

V tejto kapitole budú zároveň popísané knižnice jazyka Python ako Darts, Pandas, Plotly či Statsmodels, ktoré sa v rámci implementácie využili na vykonávanie úloh ako je predpovedanie, štatistická analýza či vizualizácia. Každá použitá knižnica bude predstavená, priblížia sa jej kľúčové funkcie, ako aj využitie týchto knižníc v práci.

5.1 Návrh procesu analýzy

Proces analýzy časových radov zahŕňa niekoľko kľúčových krokov, ktoré je nutné vykonať zvyčajne vo vopred určenom poradí, preto je tento postup možné znázorniť pomocou diagramu. Návrh postupu práce pri analýze časových radov, zobrazuje diagram na obrázku 5.1. V hornej časti diagramu sa nachádza prvý krok, ktorým je načítanie dát. Tento krok sa môže líšiť v závislosti od konkrétneho zdroja, z ktorého budú údaje načítané, rovnako ako od prostredia, ktoré sa bude pre prácu využívať. Všetky dáta použité v tejto práci

[1] https://www.python.org/
[2] https://jupyter.org/

sú uložené vo formáte CSV, preto je pre načítanie možné použiť Python knižnicu Pandas, ktorá načíta dáta a vytvorí z nich dátový objekt, tzv. `DataFrame`. Po úspešnom načítaní údajov nasleduje krok predspracovania a úpravy dát. Tento krok zahŕňa postupy ako čistenie a transformáciu dát, doplnenie chýbajúcich hodnôt, proces nazývajúci sa vyhladenie dát (snaží sa o odstránenie výkyvov a šumu, aby dáta bolo možné ľahšie analyzovať), rozdelenie dátovej sady na časť určenú pre trénovanie a testovanie, ale aj úpravu dát do formátu, ktorý požadujú jednotlivé algoritmy.

Po dokončení úvodných krokov načítania a spracovania dát, je možné prejsť na vytváranie vizuálnych reprezentácií dát, teda grafov. Tieto grafy slúžia pre prvotnú vizuálnu analýzu, ktorá môže pomôcť lepšie pochopiť a získať obraz o vzorcoch a trendoch v údajoch. V niektorých prípadoch, hlavne ak charakter dát nie je známy, môže byť užitočné vizualizovať dáta pred ich úpravou, aby sa tak získal základný prehľad o vlastnostiach časového radu. V prípade, že sú dáta nepravidelné, obsahujú výraznú mieru fluktuácie alebo v sade chýbajú hodnoty, je dobrým krokom dáta najskôr predspracovať. Taktiež môžu nastať prípady v ktorých je potrebné dáta predspracovať pred a aj po vizualizácii.

Ak osoba vykonávajúca vizuálnu analýzu, nevie z grafu odhadnúť charakter a vzory časového radu, prípadne si nie je istá tým, že časový rad spĺňa podmienku stacionárnosti, a jeho nasledujúci postup ju vyžaduje, je vhodné pridať krok, ktorý zahŕňa sofistikovanejšie štatistické metódy pre overovanie stacionárnosti (prítomnosti jednotkového koreňa) akou je napríklad ADF test popísaný v kapitole 2.5.3.

Pre lepšie porozumenie vzorcov a korelácií dát časových radov, je dôležitým ďalším krokom analýza grafov autokorelačnej a parciálne autokorelačnej funkcie, ktoré sú základným stavebným kameňom pri analýze. Tieto grafy môžu pomôcť odhaliť skryté vzory, pomôcť pri výbere modelu, ale aj pri stanovovaní jeho parametrov.

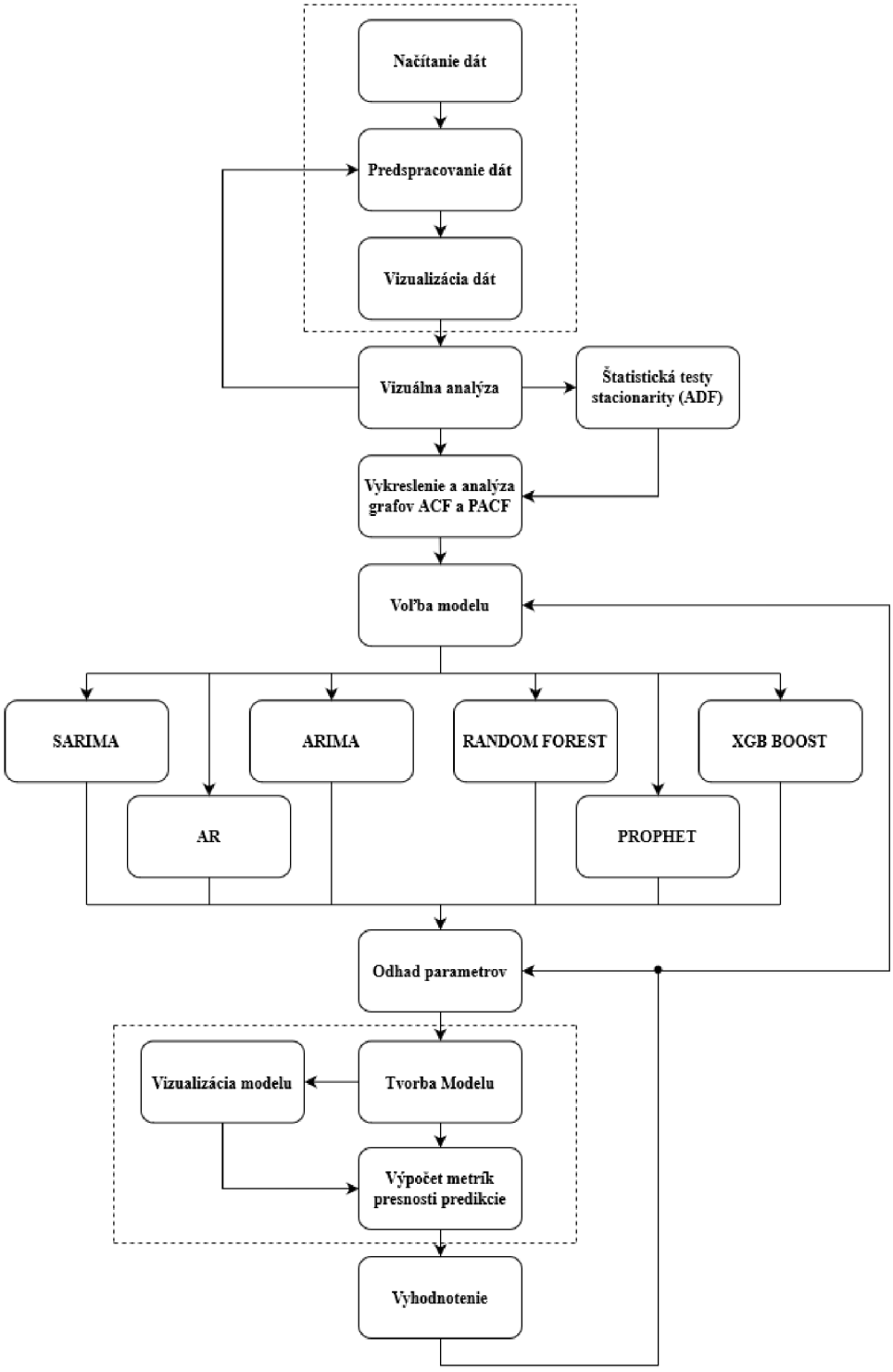
Výber správneho modelu je nevyhnutný pre spoľahlivú a hlavne presnú analýzu, jedná sa však o veľmi zložitý a citlivý proces, ktorý si vyžaduje starostlivé zváženie jednotlivých faktorov časového radu, akými sú charakter údajov, trend či sezónnosť. Z grafu 5.1, je patrná možnosť výberu niekoľkých algoritmov ktoré boli pre túto prácu zvolené. Jedná sa o základné algoritmy pre analýzu časových radov, ktorými sú AR-X, ARIMA či SARIMA, ale aj sofistikovanejšie algoritmy strojového učenia ako Random Forest, Facebook Prophet alebo XGB Boost.

Po zvolení konkrétneho modelu, ktorý sa použije pre modelovanie, je ďalším dôležitým krokom čo najpresnejší odhad jeho parametrov. Tento proces je kritickým krokom, nakoľko zahŕňa identifikáciu najlepších hodnôt parametrov modelu, ktoré pomôžu pri presnom modelovaní základných vzorov v údajoch. Celkovo sa považuje za jednu z najnáročnejších častí celej analýzy a vyžaduje dobré pochopenie základných charakteristík časového radu, používaných modelov, ale aj spoľahlivé štatistické a matematické zručnosti na presné vyhodnotenie a optimalizáciu modelov.

Nasleduje tvorba modelu na základe identifikovaných parametrov a zvoleného algoritmu. Tento krok zahŕňa prispôsobenie modelu pomocou parametrov a vytvorenenie prognózy predikcie budúcich hodnôt. V závislosti od zvoleného algoritmu sa môže jednať o vytvorenie autoregresného modelu alebo modelu strojového učenia.

Akonáhle je model vytvorený a predikcie hotové, prechádza sa k vizualizácii výsledkov a validácii modelu ktoré vykonáva sám analytik. Validácia sa vykonáva na základe niekoľkých faktorov, kde prvým z nich je vizuálna kontrola presnosti daného modelu. Hlavným cieľom kontroly grafu je rozpoznanie, či daná predikcia vôbec odpovedá a kopíruje očakávané charakteristiky a vzory pôvodného časového radu. Po nej nasleduje zhodnotenie algoritmu pomocou metrík slúžiacich k vyhodnoteniu výkonnosti a presnosti predpovedí,

ako aj porovnanie výsledkov získaných z rôznych algoritmov. Takýmito metrikami sú napríklad metriky ako priemerná absolútna odchýlka alebo smerodatná odchýlka. V prípade, že model nevyhovuje, je možné vykonať doladenie a úpravu parametrov, odhadnúť nové parametre alebo zvoliť úplne nový model/algoritmus.



Obr. 5.1: Graf návrhu procesu analýzy časového radu v jednotlivých krokoch. Orámovanie predstavuje kroky, ktoré môžu byť automatizované a nevyžadujú ľudské rozhodovanie

5.2 Implementácia

Primárnym zameraním tejto podkapitoly je poskytnúť podrobný prehľad o technických detailoch, jednotlivých zvolených knižniciach jazyka Python a ich využití v tejto práci. Konkrétne sa kapitola bude venovať piatim hlavným knižniciam a to pandas, statsmodel, plotly, darts a pmdarima. Táto podkapitola poskytne základný prehľad ku každej spomenutej knižnici a vysvetlenie akým spôsobom boli použité na implementáciu rôznych komponentov tejto práce.

5.2.1 Knižnica Pandas

Pandas[3] je veľmi výkonná a v dátovej analýze rozsiahle používaná open-source knižnica, určená pre manipuláciu s rôznymi dátami. Táto knižnica bola špeciálne navrhnutá tak, aby poskytovala efektívne a flexibilné dátové štruktúry, ako aj funkcie pre manipuláciu, analýzu a vizualizáciu týchto dát.

Jadro tejto knižnice stojí na dvoch primárnych dátových štruktúrach: Series a DataFrame. Trieda Series označuje jednorozmerné pole schopné uchovávať rôzne typy dát, zatiaľ čo DataFrame je dvojrozmerná dátová štruktúra tabuľkového typu, pozostávajúca z riadkov a stĺpcov. Pre tieto triedy bola vytvorená bohatá sada funkcií a metód pre manipuláciu s dátami, ako napríklad orezanie, filtrovanie, zlučovanie či agregovanie dát. Okrem manipulácie s dátami vyniká táto knižnica aj v analýze dát a ponúka širokú škálu štatistických funkcií a metód na sumarizáciu údajov a výpočet popisných štatistík.

Mimo toho knižnica pandas zjednodušuje a podporuje čitanie a zápis údajov z viacerých súborových formátov, vrátane CSV, SQL databáz, Excel-u a ďalších. Ďalšou vlastnosťou tejto knižnice je, že sa bezproblémovo integruje s inými vedeckými výpočtovými knižnicami, akými sú NumPy a Matplotlib, čo umožňuje jednoduchú manipuláciu a vizualizáciu dát.

V tejto práci sa v prevažnej miere využíva Trieda DataFrame a k nej príslušné funkcie a metódy na prácu s dátami a ich úpravu.

5.2.2 Knižnica Statsmodels

Knižnica Statsmodels[4] je komplexnou knižnicou určenou pre štatistické modelovanie, ktorá poskytuje širokú škálu funkcií, metód a modelov pre analýzu údajov, testovanie hypotéz a modelovanie rozličných javov. Táto knižnica je postavená na knižniciach NumPy a Pandas a častokrát je považovaná za základný nástroj pre štatistickú analýzu a modelovanie v jazyku Python.

Kľúčovou vlastnosťou Statsmodels je jej podpora lineárnych regresných modelov, ktorých ponúka množstvo, od modelu obyčajných najmenších štvorcov až po robustnú regresiu. Okrem lineárnej regresie zahŕňa aj komplexnú sadu modelov pre analýzu časových radov, akými sú napríklad modely ARIMA alebo SARIMA. Svoje možnosti použitia ďalej rozširuje rôznymi štatistickými testami, podporou testovania hypotéz a metódami pre diagnostiku modelov.

Hlavné použitie tejto knižnice v práci, je založené na využití dvoch algoritmov pre predikciu časových radov: autoregresného exogénneho modelu (AutoReg) a modelu SARIMA (SARIMAX). Ďalšími využitými funkciami sú ADF test pre test prítomnosti jednotkového koreňa a Holt-Wintersova exponenciálna vyhladzovacia metóda.

[3] https://pandas.pydata.org/

[4] https://www.statsmodels.org/stable/index.html

5.2.3 Knižnice určené pre vizualizáciu

Treťou dôležitou open-source knižnicou využitou pri implementácii, je knižnica určená pre vizualizáciu dát s názvom Plotly[5]. Plotly užívateľovi poskytuje všestrannú platformu pre vytváranie vysokej škály vizualizácií, vrátane vedeckých a štatistických grafov, obchodných panelov a rôznych iných. Je navrhnutá tak, aby výstupom boli vysokokvalitné vizualizácie pripravené pre publikovanie so zameraním na interaktivitu a možnosť prispôsobenia. Grafy reagujú na interakciu používateľa, akou je posúvanie, premiestňovanie, približovanie, zobrazenie hodnôt a podobne, čo umožňuje intuitívne analyzovať dáta a ich grafy.

Samozrejmosťou je, ako pri podobnom type knižníc, podpora širokej škály typov grafov vrátane bodových, čiarových, stĺpcových, koláčových ako aj tepelných máp či 3D plošných grafov. Rovnako ponúka rozsiahlu kontrolu nad vizuálnymi prvkami, akými sú farby, čiary, anotácie, štýly, značky a vlastnosti osí, čo tejto knižnici pridáva na flexibilite. Práve kvôli týmto vlastnostiam bola táto knižnica využitá pre vizualizácie väčšiny grafov.

Druhou knižnicou, lepšie povedané modulom knižnice, využitom pre úpravu vizualizácie niektorých grafov v tejto práci je modul matplotlib.plotly[6], ktorý umožňuje vytvárať rôzne typy grafov a upravovať ich pomocou niekoľkých jednoduchých príkazov.

5.2.4 Knižnica Darts

Darts[7] je knižnicou, ktorá bola špeciálne navrhnutá pre analýzu a predpovedanie časových radov s cieľom poskytnúť flexibilný a jednoducho použiteľný framework, ktorý ponúka širokú škálu nástrojov, modelov a pomôcok pre podobný typ úloh. Darts je skratkou pre "Data Analytics and Reporting Toolkit for Time Series", čo je možné preložiť ako súbor nástrojov pre analýzu a vytváranie prehľadov časových radov.

Najzaujímavejšou vlastnosťou knižnice Darts je jej rozsiahla zbierka modelov, zahŕňajúca modely ako ARIMA, SARIMA, Prophet či LSTM, ktoré pokrývajú široké spektrum prognostických techník a umožňujú používateľom vybrať si najvhodnejší prístup zo širokej ponuky na základe ich špecifických dát. Rovnako sa snaží o jednotnosť modelov, to znamená, že si používatelia môžu na svoje dáta aplikovať rôzne modely, porovnávať ich výkon a vybrať najlepší model na základe rôznych metrík vyhodnocovania (ktorých rovnako ponúka množstvo).

Okrem prognostických modelov a metrík ponúka komplexnú sadu nástrojov pre manipuláciu a predspracovanie časových radov, ako je doplnenie chýbajúcich hodnôt, prevzorkovanie, škálovanie, diferencovanie a rozklad podľa sezónnosti.

Bližšie táto práca využíva kombináciu prognostických modelov ARIMA, ktorý slúži pre oba algoritmy ARIMA ako aj SARIMA, RandomForest, Prophet a XGBModel. Ďalšími použitými komponentami sú funkcie pre výpočet a vykreslenie grafu funkcií ACF a PACF, funkcia check_seasonality, kontrolujúca sezónne vzory vrámci časového radu a trieda TimeSeries, ktorú tieto modely a funkcie knižnice využívajú. Ďalej sú využívané štyri metriky pre validáciu a porovnávanie modelov a to MAPE (Mean Absolute Percentage Error), SMAPE (Symmetric Mean Absolute Percentage Error), RMSE (Root Mean Squared Error) a MAE (Mean Absolute Error).

[5] https://plotly.com/python/
[6] https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html
[7] https://unit8co.github.io/darts/

5.2.5 Knižnica pmdarima

Knižnica pmdarima[8] rozširuje funkčnosť knižnice statsmodels poskytnutím dodatočných nástrojov pre automatizáciu procesu trénovania a výberu parametrov modelov ARIMA. Jej cieľom je zjednodušiť proces modelovania ARIMA automatizáciou identifikácie optimálnych parametrov a poskytuje možnosť automatického výberu modelu. Implementuje algoritmus, ktorý systematicky prehľadáva priestor parametrov modelov ARIMA, aby našiel optimálnu konfiguráciu na základe rôznych štatistických kritérií ako sú AIC (Akaike Information Criterion) a BIC (Bayesian Information Criterion). Táto automatizácia odbremení používateľov od manuálneho zadávania parametrov modelu. Konkrétne práca využíva funkciu auto_arima, ktorá spomenutý automatický proces implementuje.

5.2.6 Detekcia anomálií

Dôležitou súčasťou implementácie, bolo moje navrhnutie a vytvorenie funkcie/algoritmu, pre detekciu anomálii v testovacej sade na základe predikcie. Táto funkcia má ako vstupné parametre predikované hodnoty, aktuálne hodnoty (testovaciu sadu), názov x-ovej a y-ovej osi, povolenú odchýlku alfa a veľkosť okna, z ktorého sa určuje horný a spodný prah (tento parameter bude vysvetlený v nasledujúcom texte).

Základ tohoto algoritmu bol mnou implementovaný na princípe tzv. "rolling window" (posuvného okna). Pre každú hodnotu v testovacej sade vyberie okno predikcií na základe parametru *window_size* určujúceho veľkosť okna, ktoré je centrálne umiestnené okolo aktuálneho indexu. V rámci tohto okna sa následne určí maximálna a minimálna hodnota, ktoré sa využijú pre výpočet horného a dolného prahu. Práve maximum a minimum spolu s povolenou odchýlkou alfa určujú rozsah hodnôt, ktoré sú považované za normálnu premávku. Normálne hodnoty sa teda pohybujú v intervale $(min * (1 - \alpha), max * (1 + \alpha))$ a Konštanta alfa zohľadňuje fakt, že predikcie sú zaťažené chybou.

Akonáhle sú prahy určené, algoritmus prechádza k detekcii anomálií. V prípade, že je hodnota časového radu aktuálnych (testovacích) hodnôt väčšia ako určený horný prah, príp. menšia ako prah dolný, indikuje to potencionálnu anomáliu. Aby algoritmus rozoznal skutočné anomálie od náhodných odchýlok, vyhodnotí anomáliu až v prípade, že minimálne tri po sebe idúce hodnoty prekročia prah. Ak algoritmus anomáliu detekuje, nájde začiatok a koniec anomálie a vypíše informačnú správu s presnou pozíciou na X-ovej osi, o tom, že bola nájdená anomália.

[8] https://alkaline-ml.com/pmdarima/

Kapitola 6

Experimentálne výsledky

Experimentálnu časť tejto práce je možné rozdeliť na dva tematické celky na základe ich zamerania. Prvá časť sa snaží o potvrdenie hypotézy, že pre časové rady priemyselnej premávky, ktoré boli popísane v kapitole 4, je možné vytvoriť model a tým predikovať ich vývoj a charakter v čase s určitou presnosťou. Pre túto časť boli prvotne zvolené 4 algoritmy a to AR, MA, ARIMA a SARIMA. Tieto algoritmy boli kvôli problémom, ktoré budú popísane neskôr v tejto kapitole, rozšírené o tri ďalšie algoritmy, ktorými sú Facebook Prophet, Random Forest a XGB Boost. Algoritmus MA bol odstránený a algoritmus AR bol rozšírený na exogénny model AR-X, ktorý je schopný pracovať aj so sezónnou zložkou.

Druhá časť experimentálnej činnosti spočíva v pokusoch odhaliť, pomocou predikcie a detekcie anomálií, jednotlivé útoky v časových radoch. Pre túto časť sa na základe analýzy z prvej časti, vybrali algoritmy AR-X, Prophet a XGB Boost.

6.1 Výsledky modelovania jednotlivých časových radov

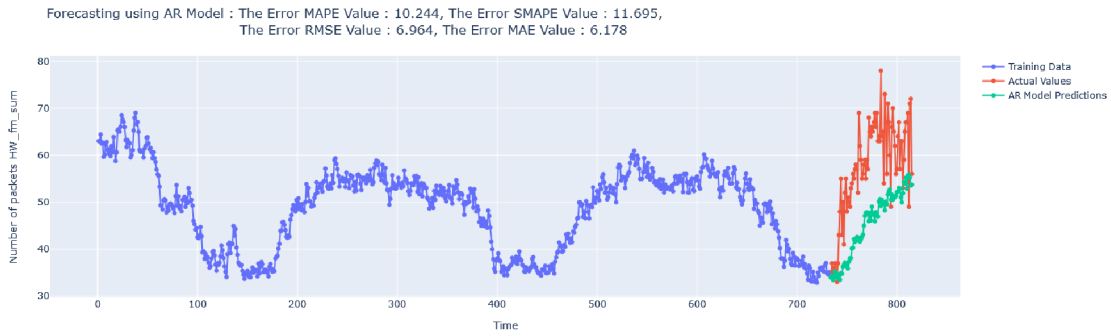
Táto časť práce bola zameraná predovšetkým na preskúmanie presnosti predikcie a výkonu algoritmov pri výbere vhodného modelu pre poskytnuté dátové sady priemyselnej premávky. V rámci každej sady prebiehala analýza a predikcia pre oba smery premávky, teda stĺpce fm_sum a tm_sum. Tieto dva stĺpce predstavujú súčet prichádzajúcich paketov k master stanici a súčet odchádzajúcich paketov z master stanice, čím poskytujú dôležité informácie o toku premávky.

V rámci tejto časti sa vyhodnocovala presnosť predikcie a výkon jednotlivých algoritmov. Pre tento účel bolo vybraných niekoľko metrík slúžiacich pre vyhodnotenie prediktívneho výkonu jednotlivých algoritmov, zahŕňajúc strednú absolútnu percentuálnu odchýlku (MAPE), sezónnu strednú absolútnu percentuálnu odchýlku (SMAPE), druhú odmocninu strednej kvadratickej odchýlky (RMSE), strednú absolútnu odchýlku (MAE) a samozrejme vizuálne vyhodnotenie časových radov. Grafy vizualizácie výsledkov predikcií všetkých algoritmov pre každú časovú radu sa nachádzajú v prílohe B. Použitie takejto kombinácie viacerých hodnotiacich metrík, poskytne komplexnejšie a dôveryhodnejšie vyhodnotenie presnosti predikcií. Cieľom je objektívne zhodnotiť presnosť a výkonnosť každého algoritmu, kde sa následne výsledky porovnajú, pričom sa zohľadnia obmedzenia každého modelu.

6.1.1 Zhodnotenie algoritmov

Pre stĺpce fm_sum dosiahol model AR-X, spomedzi použitých algoritmov, najnižšie hodnoty hodnotiacich metrík pre všetky štyri časové rady. Vypočítané metriky jednotlivých

algoritmov predikcie dátovej sady normal-traffic sú uvedené v tabuľke 6.1, ďalšie tabuľky pre zvyšné časové rady sa nachádzajú v prílohe A tejto práce. Konkrétne hodnoty metrík algoritmu AR-X pre dátovú sadu normal-traffic sú: MAPE (10.244 %), SMAPE (11.695 %), RMSE (6.964) a MAE (6.178). Tento výsledok je veľmi zaujímavý s prihliadnutím na fakt, že algoritmus AR-X je najjednoduchším modelom použitým pre predikciu jednotlivých časových radov. Odchýlka v prípade dátovej sady normal-traffic, ktorú je možné vidieť na obrázku 6.1 je spôsobená nielen chybou predikcie, ale aj mierne odlišným charakterom sady použitej na trénovanie a testovanie, kde sa na konci testovacej sady nachádza mierny nárast hodnôt oproti predchádzajúcemu sezónnemu vzoru. Práve prítomnosť sezónnej zložky viedla k nepriaznivým výsledkom algoritmu ARIMA a jedným z najvyšších hodnôt metrík pre všetky časové rady. To bolo spôsobené práve obmedzením tohto algoritmu modelovať časové rady obsahujúce silnú sezónnu zložku (s rovnakým problémom sa potýkal model MA, preto bol počas práce nahradený novými algoritmami).



Obr. 6.1: Graf predikcie modelu AR-X vyhladeného časového radu fm_sum pre dátovú sadu normal-traffic

Nečakane sa vyskytli problémy aj pri algoritme SARIMA, ktorý by v teórii nemal mať problém s podobnými radmi. Avšak, nakoniec bolo použitie práve tohto algoritmu pri všetkých analyzovaných časových radoch najproblematickejšie. Aj keď by tento algoritmus mal zvládať časové rady s rôznou dĺžkou periódy sezónnej zložky, v implementácii použitých knižníc algoritmus SARIMA nie je dostatočne optimalizovaný pre tak vysoké periódy, aké obsahujú tieto časové rady (288 okien). Pri použití tohto algoritmu s takou vysokou periódou nepomohlo ani vyskúšanie iných knižníc a predikcia sa nedokončila ani v 12-hodinovom časovom okne, čo v prípade priemyselnej premávky, kde je potrebné aby predikcie boli rýchle a dokázali sa rýchlo adaptovať na zmeny, robí tento algoritmus nepoužiteľným. S prihliadnutím na tento problém bolo nutné použiť menšiu periódu sezónnej zložky než aké v skutočnosti časové rady obsahujú (ako príklad je možné uviesť časový rad normal-traffic, pre ktorú bola zvolená perióda 12 na základe výsledkov funkcie check_seasonality popísanej v kapitole 5.2.4, podobne sa postupovalo aj v ostatných časových radoch). V dôsledku toho, aj keď sa výsledky metrík môžu javiť ako pomerne priaznivé, nedajú sa výsledky predikcie týmto algoritmom považovať za tak presné a dôveryhodné ako výsledky z iných algoritmov.

	Model	MAPE(%)	SMAPE(%)	RMSE	MAE
FM_SUM	AR-X	10.244	11.695	6.964	6.178
	ARIMA	36.71	46.44	24.592	22.434
	SARIMA	17.967	18.799	12.344	10.472
	Random Forest	19.163	21.687	13.26	11.672
	Prophet	18.105	20.38	12.634	10.984
	XGB Boost	19.56	22.284	13.564	11.826

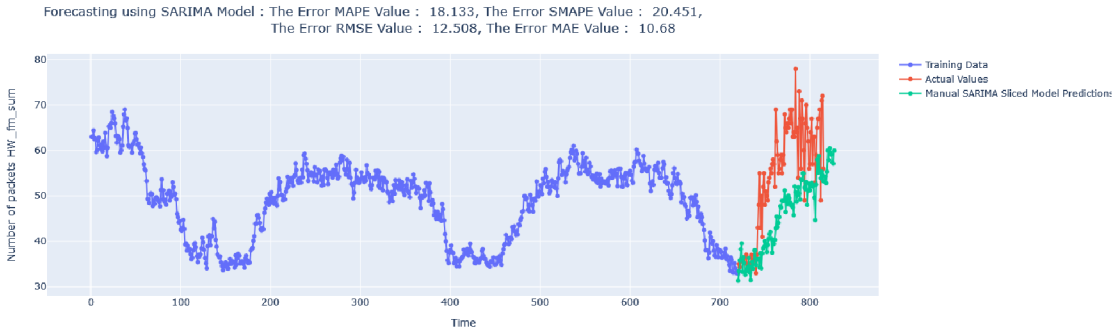
Tabuľka 6.1: Výsledky metrík slúžiace k hodnoteniu jednotlivých algoritmov pre časové rady fm_sum zo sady main normal traffic

Aby táto práca nepriniesla len záver, že algoritmus SARIMA je pre tieto dátové sady nepoužiteľný, venuje sa aj návrhu riešenia problému vysokej periódy ich sezónnej zložky. Toto experimentálne riešenie bolo implementované pre časový rad normal-traffic a spočíva v rozdelení časového radu na 12 menších časových radov, s nižšou periódou (konkrétne 24), kde každý nový časový rad vznikne priradením hodnôt z pôvodnej časovej rady s rozostupom 12 hodnôt. Pre každý takto novovytvorený časový rad sa následne prevedie analýza, odhad parametrov ako aj predikcia nových hodnôt. Pre tento postup boli zvolené dva prístupy: automatický prístup pomocou algoritmu *auto_arima*, ktorý dokáže samostatne odhadnúť najlepší model pre daný časový rad a manuálny prístup voľby parametrov. Manuálny prístup spočíval v odhadnutí parametrov pre prvý novovzniknutý časový rad s prihliadnutím na rad pôvodný, kde sa tieto parametre následne použili pre vytvorenie modelov zvyšných parciálnych časových radov. Je pravdepodobné, že manuálnym prístupom by sa dosiahli ešte lepšie výsledky, v prípade prevedenia analýzy pre každú čiastkovú radu samostatne, no takýto prístup by bol veľmi pracný, časovo oveľa náročnejší a ponúkal by väčší priestor pre zavedenie chyby do predikcií. Grafickú vizualizáciu výsledku predikcie manuálneho prístupu, je možné vidieť na obrázku 6.2. Po vytvorení predikcií pre každú čiastkovú radu, sa predikcie následne spoja do jednej finálnej časovej rady, ktorá je považovaná za finálnu predikciu. Tento prístup vyriešil problém s časovou náročnosťou vysokých períód sezónnej zložky v algoritme SARIMA a priniesol pomerne priaznivé predikcie, ktorých hodnotiace metriky sú uvedené v tabuľke 6.2.

Prístup	MAPE(%)	SMAPE(%)	RMSE	MAE
Automatický	20.023	23.027	14.294	12.036
Manuálny	18.133	20.451	12.508	10.68

Tabuľka 6.2: Výsledky metrík predikcie algoritmu SARIMA s využitím prístupu rozdelenia časového radu na 12 parciálnych časových radov

S prihliadnutím na problematiku niektorých algoritmov, boli v priebehu práce pridané tri ďalšie algoritmy, ktoré sú v porovnaní s pôvodným výberom sofistikovanejšie. Konkrétne sa jedná o algoritmy založené na princípe strojového učenia a to Random Forest, ktorý kombinuje viacero individuálnych rozhodovacích stromov pre vytváranie predikcií, Facebook Prophet, ktorý používa princíp rozkladu časového radu na jeho základné komponenty a XGB Boost založený na princípe tzv. “gradient boosting”, s princípom postupného trénovania súboru slabých predikčných modelov a každý nasledujúci model je trénovaný tak, aby opravoval chyby predchádzajúcich modelov. Podmienkou algoritmov Prophet a XGB bolo

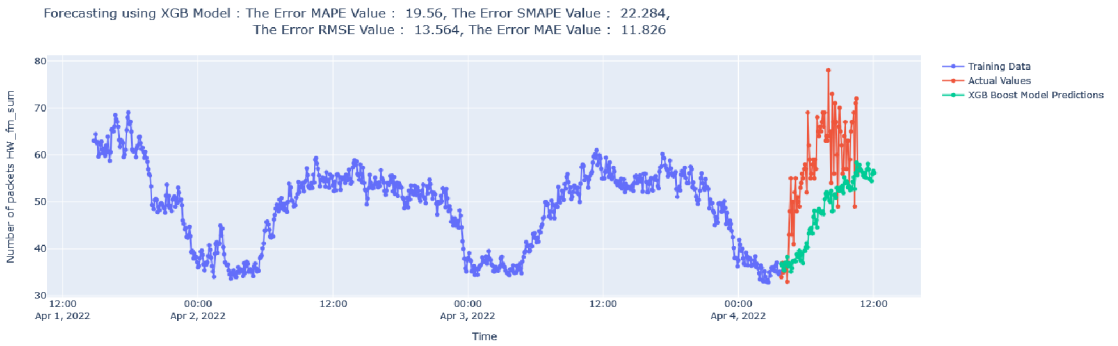


Obr. 6.2: Graf predikcie modelu SARIMA s využitím parciálnych časových radov fm_sum pre dátovú sadu normal-traffic

upravenie dátovej sady tak, aby obsahovala časovú líniu pre jednotlivé hodnoty, nakoľko tieto algoritmy berú v úvahu aj čas.

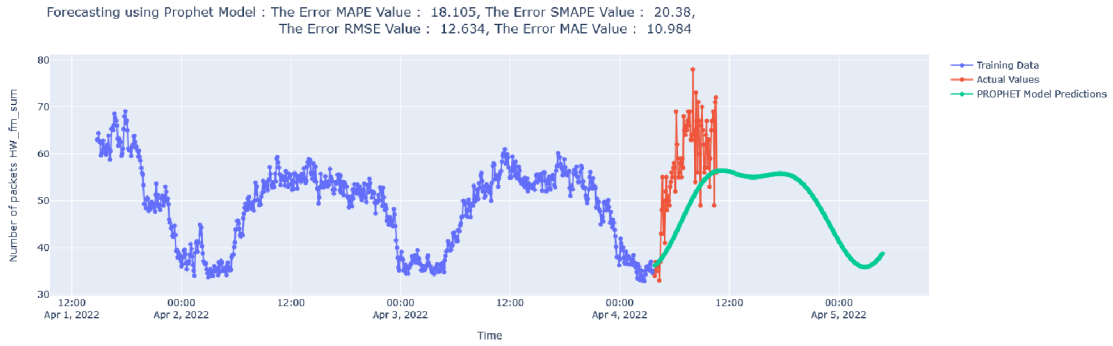
Tieto neskôr pridané algoritmy si naprieč dátovými sadami s časovou radou fm_sum počínali veľmi podobne a rozdiely vo výsledných metrikách neboli nijak výrazné. Ich predikcie pre všetky dátové sady, je možné na základe spočítaných metrík zhodnotiť v porovnaní s modelom AR-X ako veľmi dobré, a tieto algoritmy by sa dali na základe metrík spoločne zaradiť na druhú priečku v precíznosti a výkonnosti. Ak sa však v porovnaní algoritmov vezme v úvahu aj vizuálna reprezentácia v podobe grafov, teda nie len výsledné hodnoty metrík. Zo subjektívneho hľadiska by sa na druhé miesto zaradil algoritmus XGB Boost, ktorého predikcia je vyobrazená na obrázku 6.3. S prehliadnutím faktu, že z týchto troch algoritmov zvyčajne dosahoval práve vyššie hodnoty porovnávacích metrík, vedel najlepšie, aj v prípade vyhladenia pôvodného časového radu, kopírovať jeho charakter (konkrétne malé výkyvy v premávke).

Na druhú stranu sa algoritmu Facebook Prophet, darilo v predikciách lepšie vystihnúť sezónny vzor. To najmä v prípade grafov ako je možné pozorovať na obrázku 6.4, kde bol tento vzor pre ľudské oko jednoduchšie pozorovateľný ako v predikciách, ktoré sa snažili predikovať aj mierne odchýlky v premávke. To si však samozrejme vypýtalo cenu v podobe až príliš vyhladených predikcií, ktoré v grafe môžu pripomínať skôr krivku prostrednej hodnoty.



Obr. 6.3: Graf predikcie modelu XGB vyhladeného časového radu fm_sum pre dátovú sadu normal-traffic

Čo sa týka algoritmu Random Forest, z hľadiska vizuálneho charakteru predikcií by sa dal zaradiť niekde medzi tieto dva algoritmy. Je však nutné podotknúť, že výsledky jednotlivých iterácií vytvárania predikcií pomocou tohto modelu, aj za použitia rovnakých parametrov, sa môžu mierne líšiť. Mierne odchýlky sú spôsobené náhodným výberom premenných pri vytváraní rozhodovacích stromov. Grafy jednotlivých predikcií algoritmov, ako aj tabuľky s porovnaním metrík pre všetky časové rady sa nachádzajú v prílohách A a B tejto práce.



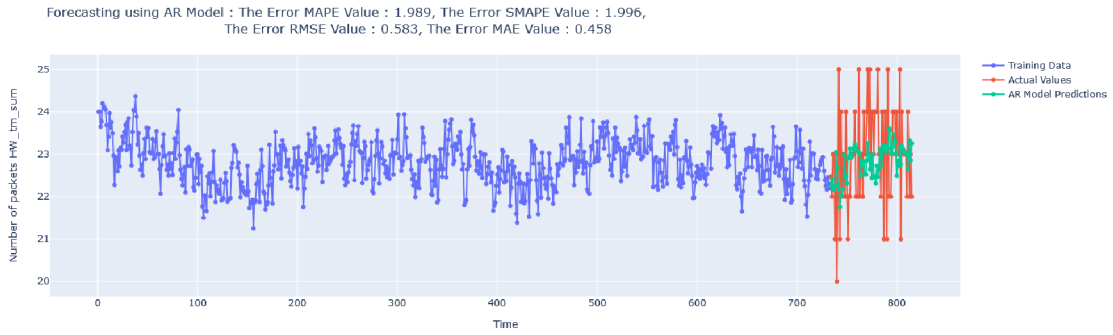
Obr. 6.4: Graf predikcie modelu PROPHET vyhladeného časového radu fm_sum pre dátovú sadu normal-traffic

Časové rady tm_sum sa naprieč dátovými sadami svojou zložitosťou a vzormi buďto podobali časovému radu fm_sum príslušnej dátovej sady, v ktorej sa nachádzali, alebo boli pre modelovanie oveľa jednoduchšie, kde práve prípadom jednoduchšej časovej rady bola dátová sada normal-traffic. Ich zložitosť bola pre predikciu nižšia, hlavne z dôvodu slabších sezónnych vzorov a väčšina z nich sa už bez úprav dala považovať za stacionárnu. S prihliadnutím na tento fakt, ako potvrdili hodnotiace metriky a vizuálna analýza vytvorených modelov, najlepšie predikcie pre všetky štyri časové rady dosiahol opäť algoritmus ARX. Kde konkrétne hodnoty metrík tohto algoritmu pre dátovú sadu normal-traffic, ako je možné vidieť z tabuľky 6.3, dosiahli nasledujúce hodnoty: MAPE (1.989 %), SMAPE (1.996 %), RMSE (0.583) a MAE (0.458). Vizualizácia tejto predikcie je vyobrazená na obrázku 6.5. Algoritmy ARIMA a SARIMA sa aj pri týchto časových radoch potýkali s podobnými problémami a nepreukázali sa ako veľmi vhodné a precíznosťou opäť zaostávali. Aj keď sa na prvý pohľad môže zdať, že sa hodnoty metrík razantne nelíšia, v niektorých sadách dokonca prekonali metrikami zvyšné algoritmy, problém týchto predikcií sa odhalil práve vo vizualizáciách, kde predikciu tvorí pomyselná, takmer rovná čiara prostrednej hodnoty z tréningovej sady.

	Model	MAPE(%)	SMAPE(%)	RMSE	MAE
TM_SUM	AR-X	1.989	1.996	0.583	0.458
	ARIMA	4.019	4.053	1.137	0.929
	SARIMA	3.944	3.97	1.123	0.911
	Random Forest	3.826	3.846	1.107	0.882
	Prophet	3.798	3.762	1.075	0.864
	XGB Boost	3.78	3.761	1.081	0.863

Tabuľka 6.3: Výsledky metrík slúžiace k hodnoteniu jednotlivých algoritmov pre časové rady tm_sum zo sady main normal traffic

Algoritmus Random Forest si pri analýze časových radov tm_sum, na základe výsledkov hodnotiacich metrík, viedol veľmi sľubne, podobne ako algoritmy PROPHET či XGB. Pri hodnotení boli zaznamenané pomerne presné predikcie, ktoré však z hľadiska vizuálnej analýzy vykazovali určitú kostrbatosť, s výnimkou dátovej sady 10days, ktorú tento algoritmus predikoval veľmi dobre. To naznačuje, že tento algoritmus nie vždy úplne presne zachytil charakter časových radov, preto nie je možné s istotou dospieť k záveru, že algoritmus poskytoval presné predikcie v ostatných dátových sadách. Vizualizácie odhaľujú rôzne výkyvy, odchýlky a prepady, ktoré narúšajú pôvodné vzory. Tieto náhodné prepady predstavujú neočakávané výkyvy, ktoré môžu byť spôsobené komplexnosťou analyzovaných dát.



Obr. 6.5: Graf predikcie modelu AR-X vyhladeného časového radu tm_sum pre dátovú sadu normal-traffic

V porovnaní s tým sa algoritmu PROPHET darilo dosahovať lepšie výsledky aj pri zohľadnení skutočnosti, že tento algoritmus poskytuje úplne vyhladené predikcie. Jeho predikcie vo všetkých dátových sadách, podobne ako pri časovej rade fm_sum, fluktuujú okolo priemernej hodnoty testovacej sady, pričom v prípade dátových sád 10days a rtu11 je možné pozorovať veľmi presnú predikciu vzoru.

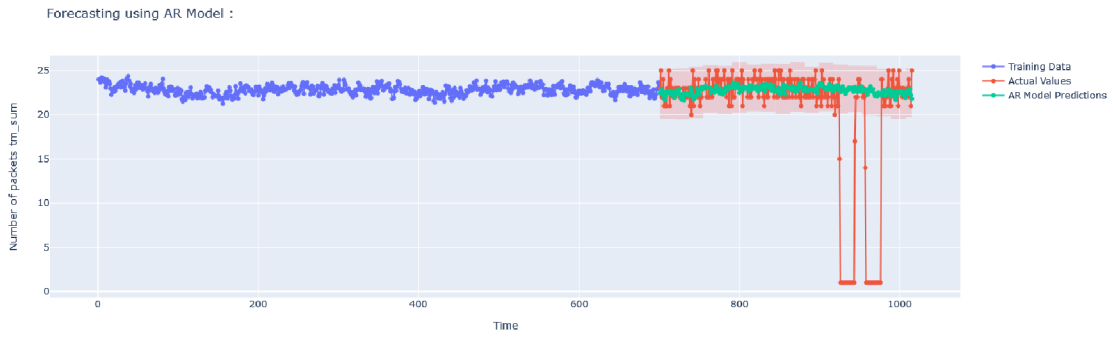
Posledným použitým algoritmom, bol algoritmus XGB Boost, ktorý si dokázal z tejto trojice asi najlepšie poradiť so všetkými časovými radmi tm_sum z poskytnutých dátových sád. Jeho predikcie dosahovali vysokú presnosť a vizuálne najvernejšie zachytávali charakter časových radov. Zaujímavosťou ktorú je možné vypozorovať z grafov, je, že aj v prípade keď došlo k miernemu vychýleniu predikcie, algoritmus sa dokázal v čase vrátiť späť a znova kopírovať vzor aktuálnych dát.

6.2 Detekcia útokov

Druhá tematická časť praktickej časti sa zaoberala detekciou anomálií a útokov v časových radoch pomocou modelu predikcie. Cieľom tejto analýzy bolo identifikovať nezvyčajné vzory a udalosti, ktoré by mohli signalizovať potencionálne útoky v sledovaných časových radoch.

Prvým krokom bolo vytvorenie prediktívnych modelov pre každý časový rad s použitím vhodného algoritmu. Na základe získaných poznatkov z prvej časti v kapitole 6.1.1, boli pre túto časť zvolené tri algoritmy AR-X, Facebook Prophet a XGB Boost. Pre úspešnú detekciu anomálií, bolo dôležité zvoliť vhodné parametre a prahové hodnoty, aby sme minimalizovali falošne pozitívne a falošne negatívne predikcie. Následne sa v rámci detekcie hľadali významné odchýlky medzi predikciou a skutočnými hodnotami časových radov.

Vo výslednom hodnotení sa časový rad tm_sum ukázal byť vhodnejším kandidátom pre detekciu útokov v priemyselnej premávke normal-traffic, než časový rad fm_sum. Je to samozrejme dôsledok toho, že práve tento rad je v danej sade tým jednoduchším časovým radom pre modelovanie. Pomocou predikcie sa tak podarilo odhaliť takmer všetky anomálie a útoky s minimálnou chybovosťou. Čo sa týka jednotlivých algoritmov, tak si v tomto prípade všetky viedli veľmi dobre. Ich výsledky so zameraním na detekciu boli vo výsledku totožné. Ako vizuálny príklad je uvedená detekcia DOS útokov za použitia algoritmu AR-X, na obrázku 6.6. Algoritmus implementovaný za účelom detekcie anomálií, ktorý bol popísaný v kapitole 5.2.6, na základe týchto predikcií a vhodne zvoleného prahu, nedetekoval žiadne falošné anomálie. Prah bol pre tento časový rad stanovený ako $\alpha = 0.1$ (teda 10 % odchýlka nahor aj nadol od určeného maxima/minima). Jediným útokom, ktorý algoritmus detekcie v žiadnom z použitých modelov nedokázal odhaliť, je, prvý útok pomocou tzv. "injekcie" čo činí 90% úspešnosť. V tomto útoku útočník namiesto klasických požiadavkov, posielal nezvyčajné požiadavky. Okrem obsahu týchto požiadavkov, sa takýto spôsob útoku neodlišuje výraznou zmenou v toku. To znamená, že sa tento útok neprejaví značným zvýšením alebo znížením počtu paketov, ktoré by sa významne odlišovali od charakteru časovej rady.



Obr. 6.6: Graf detekcie DOS útoku za použitia modelu AR-X v časovom rade tm_sum

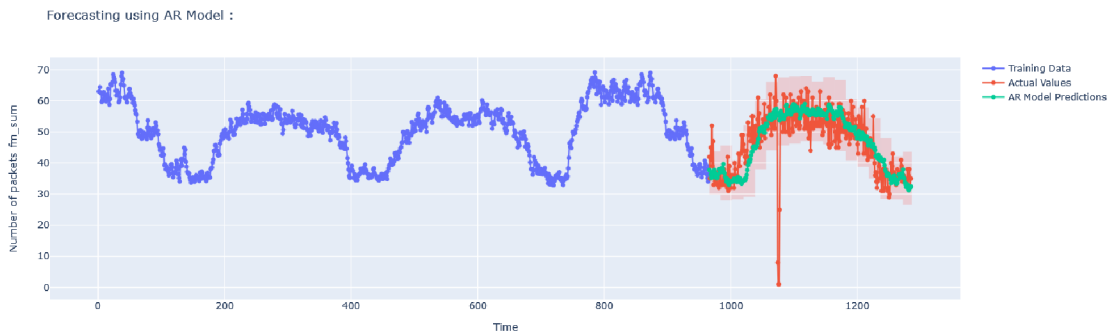
Na druhú stranu časové rady fm_sum boli v tomto ohľade problematickejšie, práve kvôli silnému sezónnemu vzoru v dátovej sade normal-traffic, ktorý so sebou priniesol rôzne problémy. Prvým problémom v týchto časových radoch (s rovnakým problémom sa potýkali aj rady tm_sum) bolo umiestnenie simulovaných útokov, kde napríklad útok typu rogue device, bol umiestnený tak, že začínal hneď vo ôsmom okne dátovej sady, tým pádom tam neexistuje priestor pre vytvorenie časti určenej pre trénovanie. Bolo tak potrebné vymyslieť

spôsob, akým vytvoriť trénovaciu a testovaciu časť, tak aby sa výrazne nenarušil charakter tejto časovej rady. Za týmto účelom bol zvolený prístup napojenia časti časového radu obsahujúceho útok, za pôvodný časový rad sady normal-traffic. Aj keď takáto úprava v radoch tm_sum nespôsobila žiadne signifikantné zmeny alebo výkyvy, v časových radoch fm_sum to prinieslo ďalšie problémy.

Prvým bol mierny nárast hodnôt v rámci jednej periódy (kde sa práve tieto zvýšené hodnoty nachádzali na začiatku a konci tejto časovej rady), čo môže mať za následok menej presnú detekciu a falošné označenie anomálií. Druhým problém, ktorý tento prístup priniesol, bolo skrátenie dĺžky jednej periódy silnej sezónnej zložky. Nakoľko hodnoty zo začiatku a konca tejto časovej rady nemali dostatočnú šírku aby skopírovali pôvodný vzor, čo malo tiež vplyv na niektoré predikcie. Ako veľmi táto skutočnosť ovplyvnila predikcie, záviselo od toho, v ktorej časti tejto rady predikcia začínala. Na základe toho bolo potrebné v niektorých prípadoch aspoň čiastočne minimalizovať tento problém zopakovaním istého úseku hodnôt. K tejto úprave však bolo potrebné pristupovať veľmi opatrne. Obmedzujúcim faktorom bolo taktiež to, že sa tento úsek nachádzal práve v oblasti spojenia konca pôvodného časového radu a začiatku radu s útokom, kde bol zaznamenaný nárast jednotlivých hodnôt prenesených paketov. Preto sa vo výsledku predĺženie pohybovalo medzi dvadsiatimi až tridsiatimi oknami.

Vzhľadom na vyššiu premenlivosť hodnôt dát a už spomenuté úpravy, počas ktorých narastala pravdepodobnosť zanesenia chyby, treba tieto predikcie vyhodnocovať s väčšou rezervou a práve kvôli tomu sa pre tieto časové rady, za účelom detekcie, určila vyššia konštanta stanovenia prahu, konkrétne $\alpha = 0.15$.

Čo sa jednotlivých algoritmov týka, najlepšie si pre tieto časové rady viedol algoritmus AR-X. Tento algoritmus, s mierne prispôsobenými parametrami na základe úprav, ktoré sa vykonali, vedel najlepšie predpovedať vývoj a vzor časovej rady. V porovnaní s radou tm_sum, sa v tomto prípade na základe predikcií, nepodarilo odhaliť ani jeden z dvoch útokov pomocou injekcie, zároveň algoritmus detekcie nebol schopný odhaliť jeden z dvoch skenovacích útokov. Dôveryhodnosť detekcie z hľadiska chybne vyhodnotených útokov sa pohybovala v priemere medzi jednou až dvomi falošne označenými anomáliami. Príklad vizualizácie a výstupu detekcie útoku skenovaním, pre predikcie časového radu fm_sum za použitia algoritmu AR-X je možné vidieť na obrázkoch 6.7 a 6.8.



Obr. 6.7: Graf detekcie útoku skenovaním za použitia modelu AR-X v časovom rade fm_sum

Predikcie algoritmov Prophet a XGB si v tomto prípade neviedli vôbec dobre. Detekcia anomálií na základe predikcií modelu Prophet, mala nižšiu mieru správne detekovaných útokov, s výrazne vyššou mierou chybovosti klasifikovania falošných anomálií z bežnej pre-

mávky. Falošné detekcie sa v niektorých prípadoch pohybovali aj v dvojciferných číslach, a s prihliadnutím na tieto skutočnosti sa algoritmus Prophet nedá považovať za vhodný pre účely detekcie anomálií. Modely algoritmu XGB sa nachádzajú svojimi výsledkami niekde uprostred, bližšie k algoritmu Prophet. Aj keď si v pri niektorých útokoch viedli veľmi dobre, pri iných bolo zaznamenané zlyhanie ako v prípade Prophet algoritmu. Bolo to podmienené hlavne miestom, v ktorom, predikcie konkrétneho časového radu začínali (podľa toho kde v rade sa útok nachádzal sa prispôsobila časť určená pre trénovanie a testovanie), ktoré algoritmus AR-X zvládal lepšie. Grafy detekcie anomálií si je možné prehliadnuť v prílohe C.

```
detect_anomaly(forecast, df_test_ar, "window_number", "fm_sum", alpha = 0.15)
```

```
Anomaly detected at window_number 1074 - 1077
```

Obr. 6.8: Použitie algoritmu pre detekciu útoku skenovaním za použitia predikcie modelu AR-X v časovom rade fm_sum

Kapitola 7

Záver

Cieľom tejto práce bolo otestovať vybrané metódy pre predikciu budúcich hodnôt časových radov so zameraním na ich využitie pri detekcii rôznych typov útokov a anomálií. Tento cieľ sa podarilo naplniť vo všetkých bodoch. Po prvotných neúspechoch počiatočných algoritmov, ktoré tvorila štvorica algoritmov AR, MA, ARIMA a SARIMA, bola práca rozšírená o ďalšie sofistikovanejšie algoritmy pomocou ktorých, sa na základe analýzy podarilo vytvoriť precízne modely pre jednotlivé časové rady. Okrem toho sa práca snažila poskytnúť návrh a otestovanie alternatívneho prístupu predikcie s využitím parciálnych časových radov, v dôsledku problematiky chýbajúcej optimalizovanosti algoritmu SARIMA pre vysoké sezónne periódy. S využitím hodnotiacich metrík a vizuálnej analýzy sa vykonalo porovnanie výkonnosti jednotlivých modelov, pričom algoritmus AR-X preukázal najlepšiu výkonnosť a konzistentne dosahoval najlepšie výsledky naprieč celým súborom časových radov. Poznatky získané z analýzy a predikcie modelov boli neskôr zúžitkované pri detekcii rôznych typov útokov. V časti zameranej na detekciu útokov sa podarilo vytvoriť algoritmus, ktorý pomocou predikcií dokázal odhaliť takmer všetky typy poskytnutých anomálií. Hlavným úspechom tejto práce, je v dokázanej schopnosti s určitou presnosťou predikovať rôzne typy premávky a využitia týchto predikcií práve pre detekciu anomálií v priemyselných sieťach.

Jedným z validných rozšírení tejto práce by v prvom rade mohlo byť rozšírenie dátovej sady o ďalšie rôznorodé typy priemyselných premávok. Zároveň by sa mohli pridať dátové sady, ktoré zaznamenávajú dlhší časový horizont, a tiež dátové sady obsahujúce reálne útoky, nie len tie nasimulované. Týmto rozšírením by sa otvorila možnosť implementovať predikčný prístup na reálnu premávku (s určitým oneskorením), v konkrétnej priemyselnej sieti. Práve takýto prístup, ako vyplynulo z výsledkov tejto práce, vykazuje potenciál pre detekciu útokov.

Ďalším rozšírením práce do budúcna, je určite rozšírenie sady algoritmov, prípadne vyskúšanie metódy učenia neurónových sietí, pre vytváranie modelov predikcie. Skvelým doplnkom by taktiež bolo vytvorenie unifikovaného automatizovaného interaktívneho prostredia pre analýzu a úpravu dát, modelovanie a vyhodnotenie predikcií, ktoré by analytika odbremenilo od veľkej časti práce, ktorú musí vykonávať ručne.

Literatúra

[1] Palachy Affek, S. *Detecting stationarity in time series data* [online]. 2019 [cit. 2023-02-22]. Dostupné z: https://towardsdatascience.com/detecting-stationarity-in-time-series-data-d29e0a21e638.

[2] Arlt, J., Arltová, M. a Rublíková, E. *Analýza ekonomických časových řad s příklady*. 1. vyd. Vysoká škola ekonomická, Fakulta informatiky a statistiky, 2002. ISBN 978-80-245-0307-3.

[3] Brockwell, P. J. a Davis, R. A. *Introduction to time series and forecasting*. 3. vyd. Springer, 2002. ISBN 978-3-319-29852-8.

[4] Cipra, T. *Analýza časových řad s aplikacemi v ekonomii*. 1. vyd. SNTL/Alfa, 1986.

[5] Cryer, J. D. a Chan, K.-S. *Time series analysis with applications in R*. 2. vyd. Spinger, 2008. ISBN 978-0-387-75958-6.

[6] DANEL, R. Predikce časové řady pomocí autoregresního modelu. *Technická univerzita Ostrava*. 2004.

[7] Dotis Georgiou, A. *Autocorrelation in Time Series Data* [online]. 2019 [cit. 2023-02-09]. Dostupné z: https://www.influxdata.com/blog/autocorrelation-in-time-series-data/.

[8] Hall, M. *Variance vs. Covariance: What's the Difference?* [online]. 2021 [cit. 2023-02-08]. Dostupné z: https://www.investopedia.com/ask/answers/041515/what-difference-between-variance-and-covariance.asp.

[9] Hančlová, J. a Tvrdý, L. Úvod do analýzy časových řad. *VŠB-TU, Ostrava*. 1. vyd. 2003.

[10] Harris, R. I. Testing for unit roots using the augmented Dickey-Fuller test: Some issues relating to the size, power and the lag structure of the test. *Economics Letters*. Elsevier. 1992, zv. 38, č. 4, s. 381–386.

[11] Hoff, J. C. *A practical guide to Box-Jenkins forecasting*. 1. vyd. Lifetime Learning Publications, 1983. ISBN 0-534-02719-9.

[12] Hyndman, R. J. a Athanasopoulos, G. *Forecasting: principles and practice*. 2. vyd. OTexts, 2018. ISBN 978-0987507112. Dostupné z: https://otexts.com/fpp2/.

[13] Hyndman, R. J. a Athanasopoulos, G. *Forecasting: principles and practice*. 3. vyd. OTexts, 2021. ISBN 978-0987507136. Dostupné z: https://otexts.com/fpp3/.

[14] Jones, J. H. *Time Series and Spectral Analysis* [online]. Standford University, 2018 [cit. 2023-3-1]. Dostupné z: https://web.stanford.edu/class/earthsys214/notes/series.html.

[15] Klufová, R., Rost, M. a Klicnarová, J. *Modelování regionálních procesů*. 1. vyd. Alfa Nakladatelství, 2012. ISBN 978-80-87197-53-0.

[16] Křivỳ, I. Analỳza časovỳch řad. *Ostravská univerzita*. 1. vyd. 2012.

[17] Mehta, S. *How to make a time series stationary?* [online]. 2022 [cit. 2023-03-05]. Dostupné z: https://analyticsindiamag.com/how-to-make-a-time-series-stationary/.

[18] Nau, R. Statistical forecasting: notes on regression and time series analysis. *Fuqua School of Business, Duke University*. 2020. Dostupné z: https://people.duke.edu/~rnau/411sdif.htm.

[19] Ozbun, A. *Unit Root In Time Series* [online]. 2021 [cit. 2023-03-04]. Dostupné z: https://medium.com/codex/unit-root-in-time-series-38d451d742ce.

[20] Paparoditis, E. a Politis, D. N. The asymptotic size and power of the augmented Dickey–Fuller test for a unit root. *Econometric Reviews*. Taylor & Francis. 2018, zv. 37, č. 9, s. 955–973.

[21] Prabhakaran, S. *Augmented Dickey Fuller Test (ADF Test) – Must Read Guide* [online]. 2019 [cit. 2023-03-02]. Dostupné z: https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/.

[22] Shin, Y. a Schmidt, P. The KPSS stationarity test as a unit root test. *Economics Letters*. 1992, zv. 38, č. 4, s. 387–392. DOI: https://doi.org/10.1016/0165-1765(92)90023-R. ISSN 0165-1765. Dostupné z: https://www.sciencedirect.com/science/article/pii/016517659290023R.

[23] Standards, N. I. of a Technology. *NIST/SEMATECH e-Handbook of Statistical Methods* [online]. U.S. Department of Commerce, 2012 [cit. 2023-3-2]. Dostupné z: https://www.itl.nist.gov/div898/handbook/.

[24] Contributor, T. *Statistical mean, median, mode and range* [online]. 2023 [cit. 2023-02-08]. Dostupné z: https://www.techtarget.com/searchdatacenter/definition/statistical-mean-median-mode-and-range.

Príloha A

Tabuľky výsledkov hodnotiacich metrík pre jednotlivé časové rady

	Model	MAPE(%)	SMAPE(%)	RMSE	MAE
FM_SUM	AR-X	10.244	11.695	6.964	6.178
	ARIMA	36.71	46.44	24.592	22.434
	SARIMA	17.967	18.799	12.344	10.472
	Random Forest	19.163	21.687	13.26	11.672
	Prophet	18.105	20.38	12.634	10.984
	XGB Boost	19.56	22.284	13.564	11.826
TM_SUM	AR-X	1.989	1.996	0.583	0.458
	ARIMA	4.019	4.053	1.137	0.929
	SARIMA	3.944	3.97	1.123	0.911
	Random Forest	3.826	3.846	1.107	0.882
	Prophet	3.798	3.762	1.075	0.864
	XGB Boost	3.78	3.761	1.081	0.863

Tabuľka A.1: Výsledky metrík slúžiace k hodnoteniu jednotlivých algoritmov pre časové rady fm_sum a tm_sum zo sady main normal traffic

	Model	MAPE(%)	SMAPE(%)	RMSE	MAE
FM_SUM	AR-X	4.314	4.242	47.301	36.782
	ARIMA	8.169	7.958	89.595	68.559
	SARIMA	10.253	9.573	106.297	83.401
	Random Forest	9.154	8.688	94.29	75.392
	Prophet	8.896	8.518	93.462	73.59
	XGB Boost	8.869	8.666	95.925	74.755
TM_SUM	AR-X	4.324	4.262	5.713	4.549
	ARIMA	8.136	7.979	9.982	7.638
	SARIMA	8.263	8.021	10.031	7.68
	Random Forest	8.516	8.459	10.379	8.079
	Prophet	8.526	8.357	10.153	7.994
	XGB Boost	9.295	9.104	11.027	8.724

Tabuľka A.2: Výsledky metrík slúžiace k hodnoteniu jednotlivých algoritmov pre časové rady fm_sum a tm_sum zo sady rtu11

	Model	MAPE(%)	SMAPE(%)	RMSE	MAE
FM_SUM	AR-X	0.435	0.434	9.1	5.652
	ARIMA	1.254	1.244	22.664	16.224
	SARIMA	1.29	1.278	23.152	16.679
	Random Forest	0.763	0.762	16.878	9.942
	Prophet	0.702	0.699	16.027	9.115
	XGB Boost	0.785	0.781	16.702	10.189
TM_SUM	AR-X	0.409	0.407	2.619	1.624
	ARIMA	0.609	0.606	4.645	2.415
	SARIMA	0.622	0.619	4.655	2.468
	Random Forest	0.643	0.64	4.74	2.552
	Prophet	0.695	0.692	4.828	2.757
	XGB Boost	0.599	0.596	4.735	2.376

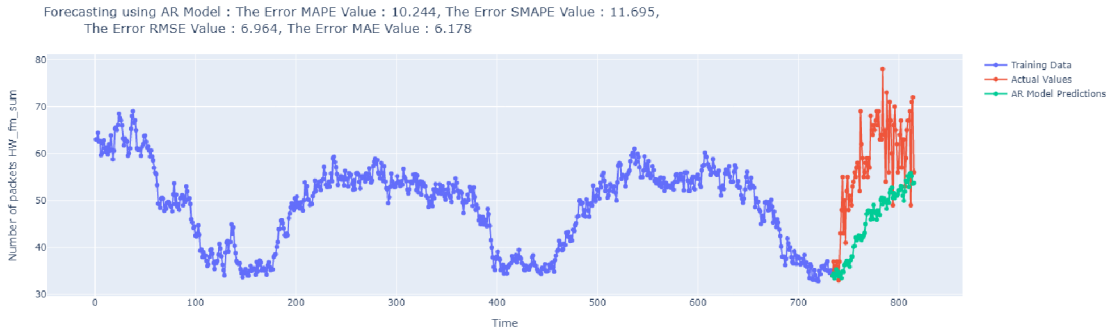
Tabuľka A.3: Výsledky metrík slúžiace k hodnoteniu jednotlivých algoritmov pre časové rady fm_sum a tm_sum zo sady 104Mega

	Model	MAPE(%)	SMAPE(%)	RMSE	MAE
FM_SUM	AR-X	1.752	1.744	4.642	3.633
	ARIMA	4.783	4.823	12.955	10.046
	SARIMA	5.025	4.969	12.935	10.353
	Random Forest	3.568	3.549	9.555	7.395
	Prophet	3.731	3.728	9.846	7.758
	XGB Boost	3.682	3.656	9.801	7.624
TM_SUM	AR-X	1.513	1.508	0.552	0.437
	ARIMA	4.068	4.031	1.444	1.169
	SARIMA	4.045	4.012	1.444	1.163
	Random Forest	3.074	3.061	1.137	0.888
	Prophet	3.088	3.081	1.133	0.893
	XGB Boost	3.251	3.235	1.185	0.938

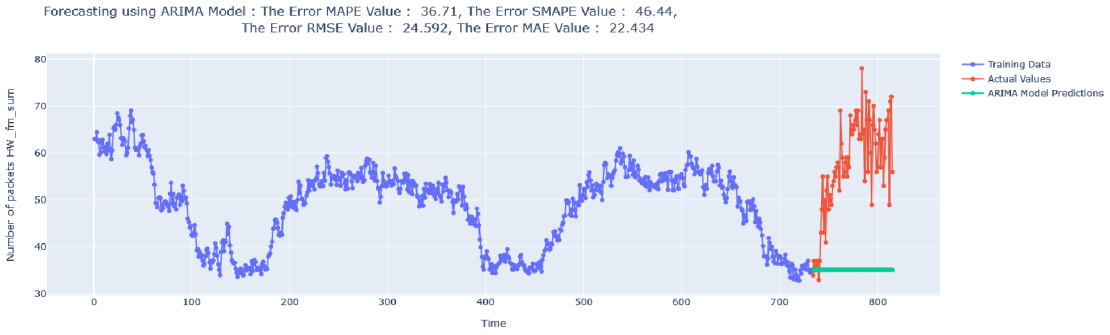
Tabuľka A.4: Výsledky metrík slúžiace k hodnoteniu jednotlivých algoritmov pre časové rady fm_sum a tm_sum zo sady 10 days

Príloha B

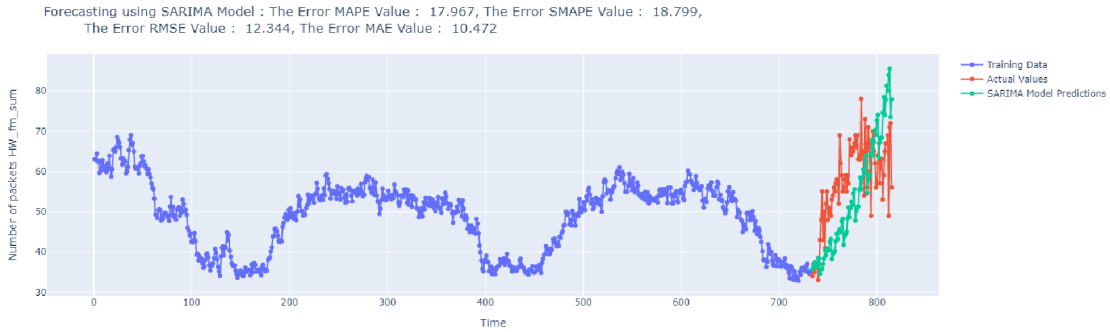
Vizualizácie jednotlivých predikcií



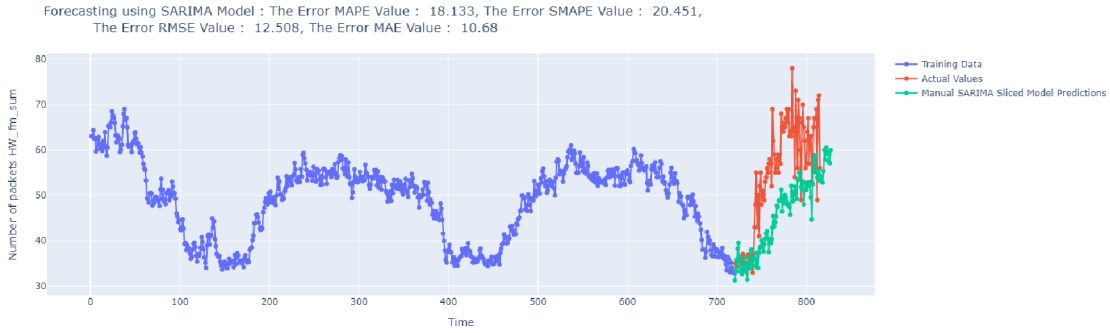
Obr. B.1: Graf predikcie modelu AR-X časového radu fm_sum pre dátovú sadu normal-traffic



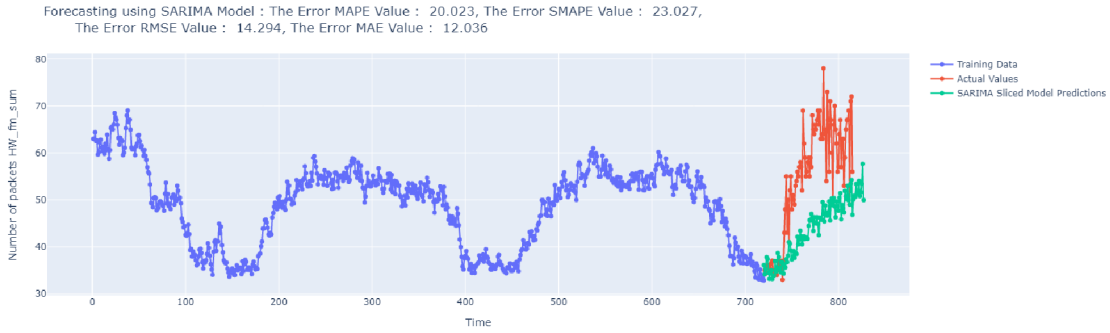
Obr. B.2: Graf predikcie modelu ARIMA časového radu fm_sum pre dátovú sadu normal-traffic



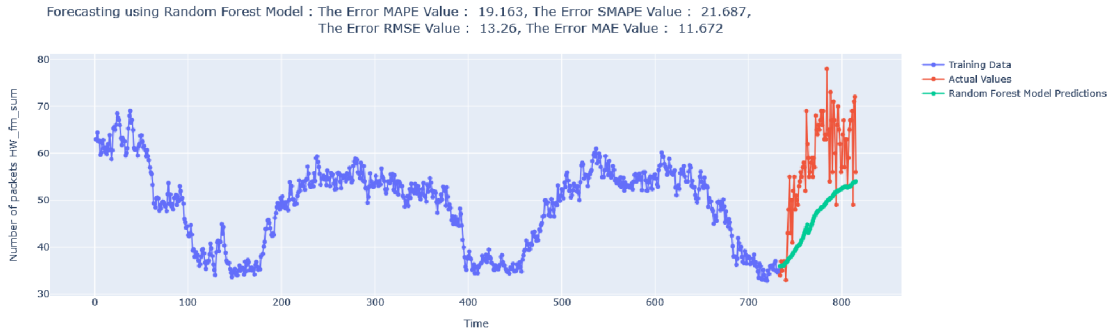
Obr. B.3: Graf predikcie modelu SARIMA časového radu fm_sum pre dátovú sadu normal-traffic



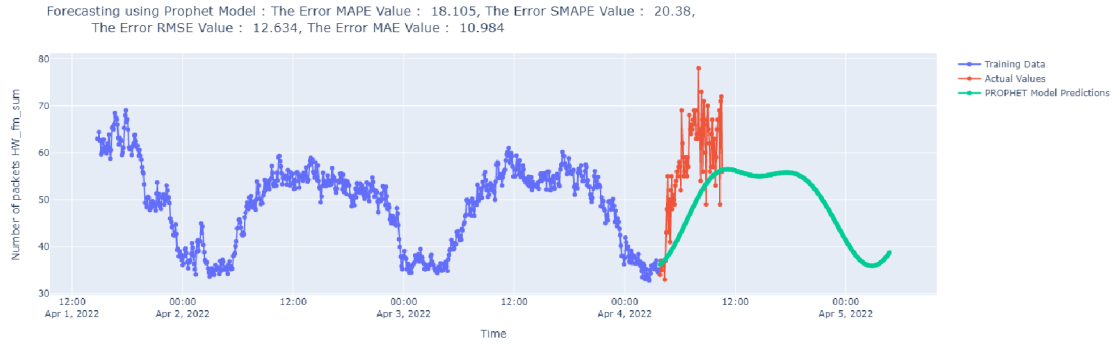
Obr. B.4: Graf predikcie modelu SARIMA s využitím parciálnych časových radov radu fm_sum s manuálnym prístupom pre dátovú sadu normal-traffic



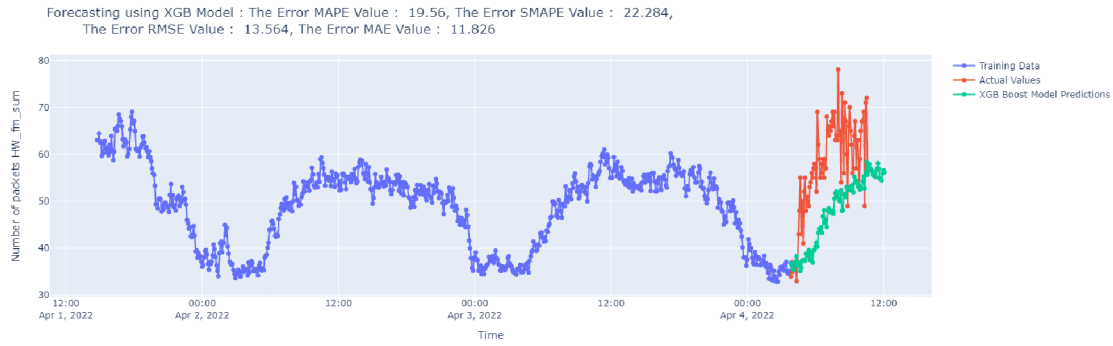
Obr. B.5: Graf predikcie modelu SARIMA s využitím parciálnych časových radov radu fm_sum s automatickým prístupom pre dátovú sadu normal-traffic



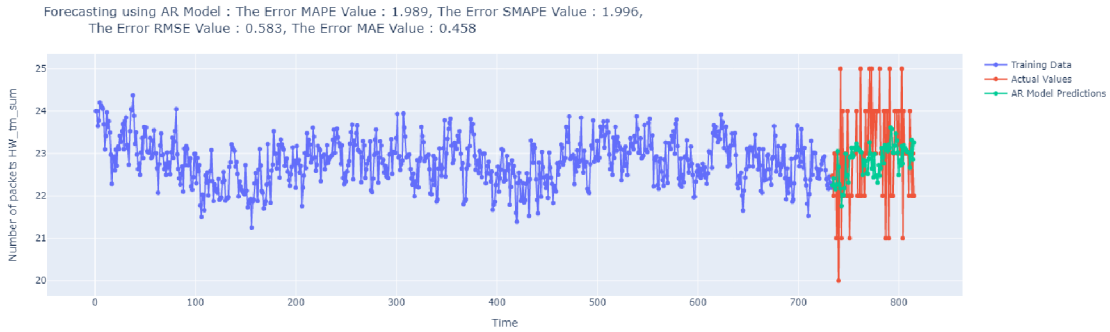
Obr. B.6: Graf predikcie modelu Random Forest časového radu fm_sum pre dátovú sadu normal-traffic



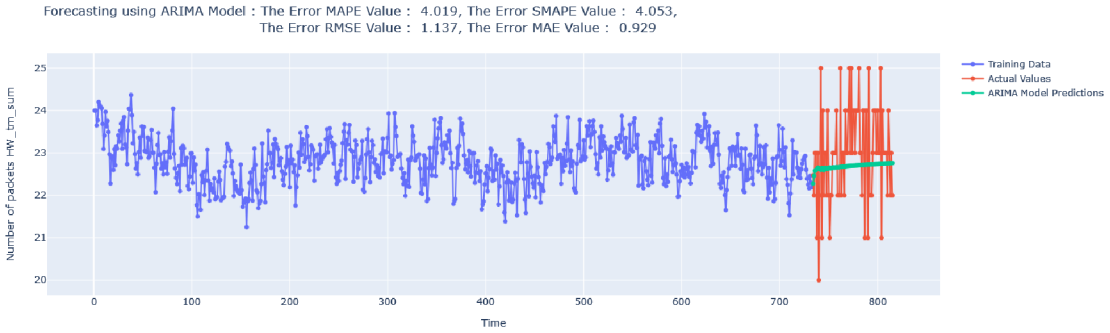
Obr. B.7: Graf predikcie modelu PROPHET časového radu fm_sum pre dátovú sadu normal-traffic



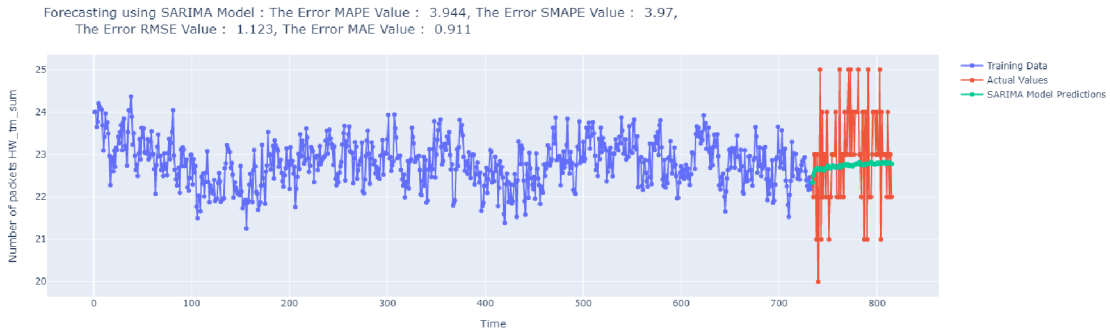
Obr. B.8: Graf predikcie modelu XGB časového radu fm_sum pre dátovú sadu normal-traffic



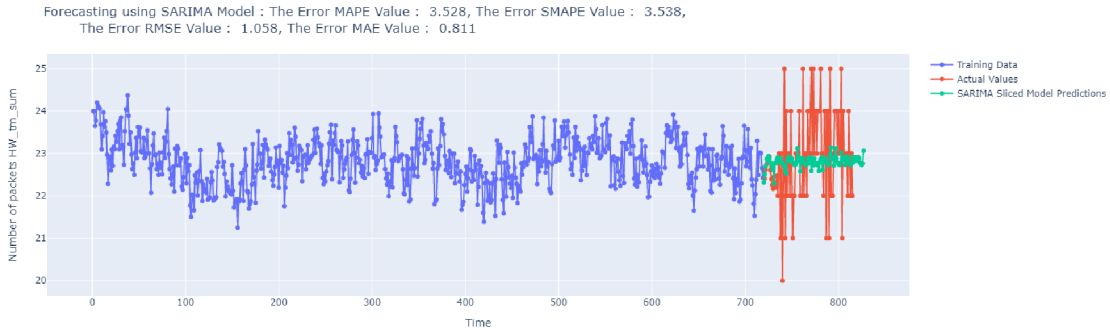
Obr. B.9: Graf predikcie modelu AR-X časového radu tm_sum pre dátovú sadu normal-traffic



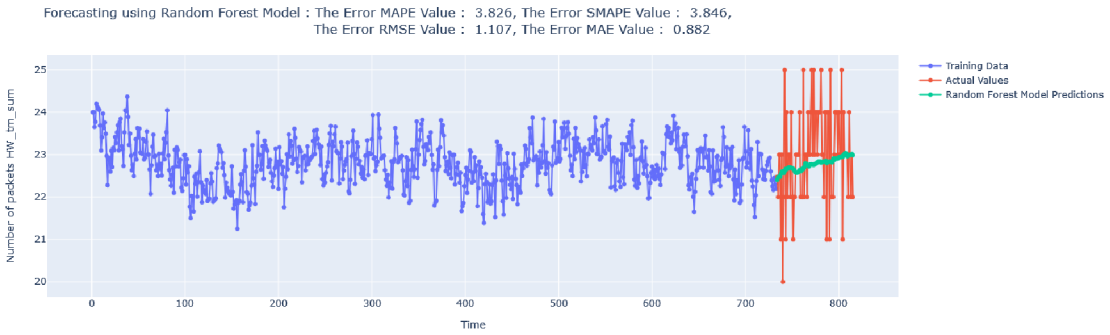
Obr. B.10: Graf predikcie modelu ARIMA časového radu tm_sum pre dátovú sadu normal-traffic



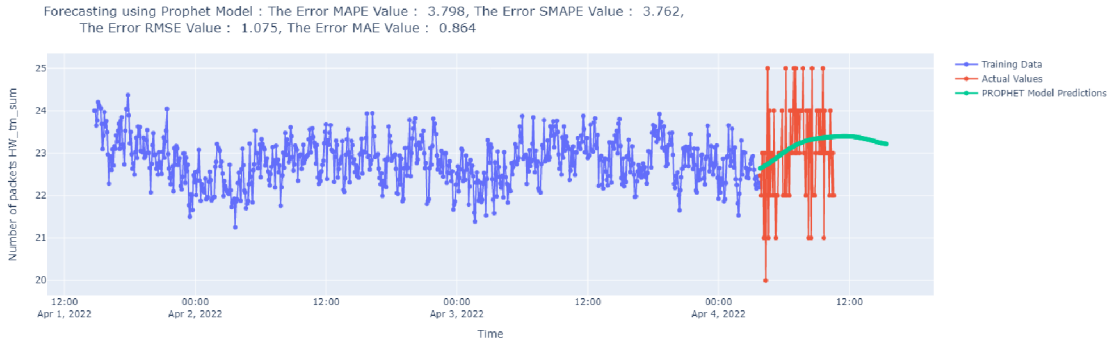
Obr. B.11: Graf predikcie modelu SARIMA časového radu tm_sum pre dátovú sadu normal-traffic



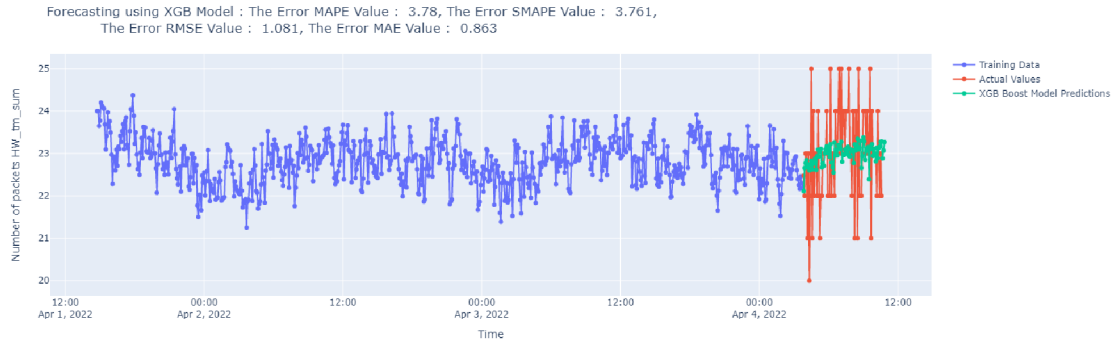
Obr. B.12: Graf predikcie modelu SARIMA s využitím parciálnych časových radov radu tm_sum s automatickým prístupom pre dátovú sadu normal-traffic



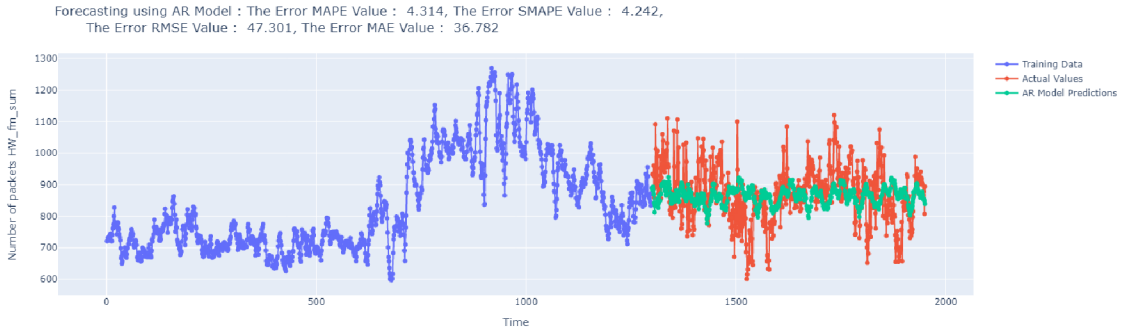
Obr. B.13: Graf predikcie modelu Random Forest časového radu tm_sum pre dátovú sadu normal-traffic



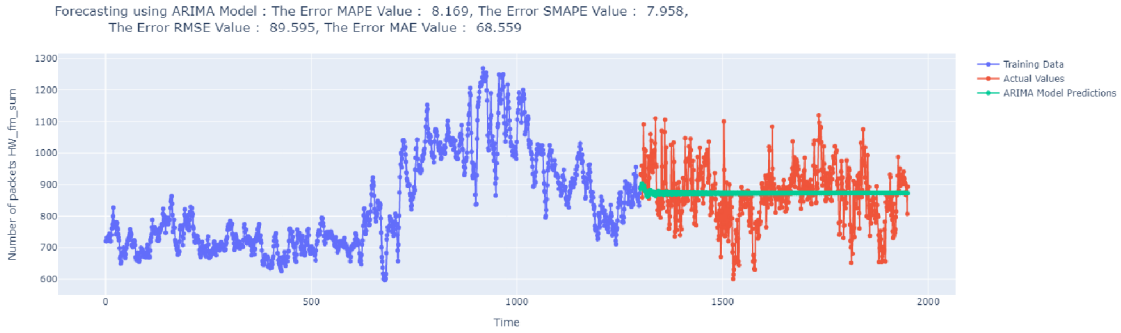
Obr. B.14: Graf predikcie modelu PROPHET časového radu tm_sum pre dátovú sadu normal-traffic



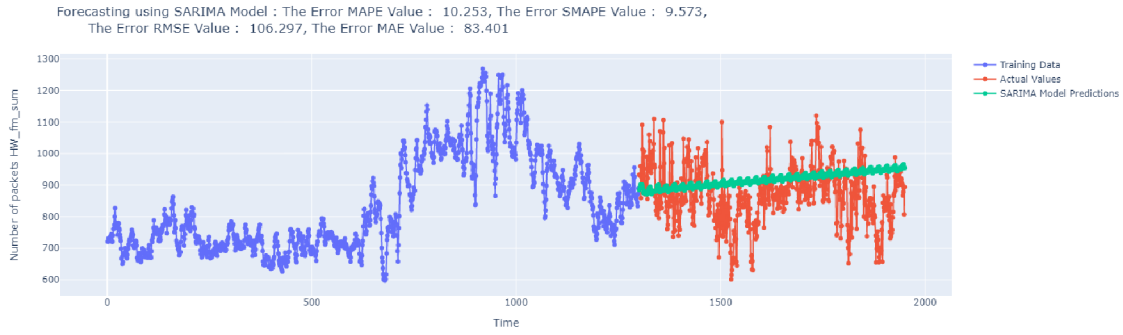
Obr. B.15: Graf predikcie modelu XGB časového radu tm_sum pre dátovú sadu normal-traffic



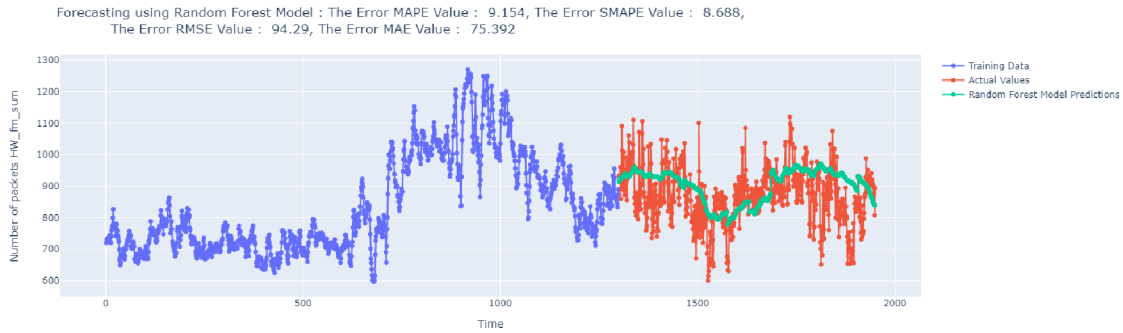
Obr. B.16: Graf predikcie modelu AR-X časového radu fm_sum pre dátovú sadu rtu11



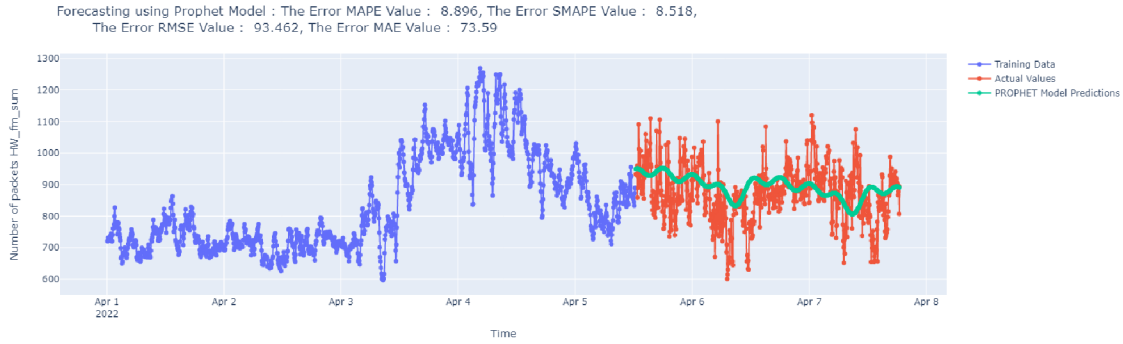
Obr. B.17: Graf predikcie modelu ARIMA časového radu fm_sum pre dátovú sadu rtu11



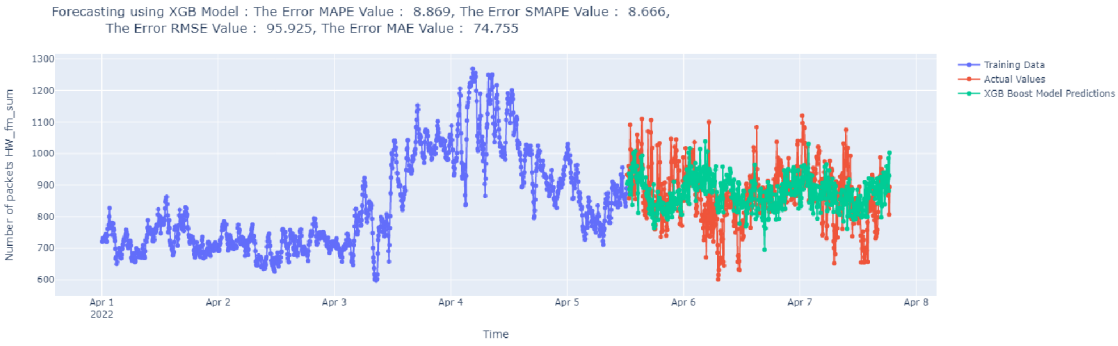
Obr. B.18: Graf predikcie modelu SARIMA časového radu fm_sum pre dátovú sadu rtu11



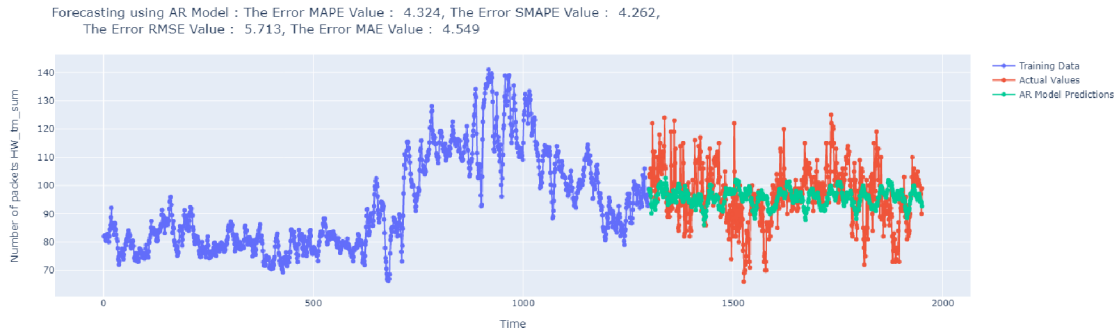
Obr. B.19: Graf predikcie modelu Random Forest časového radu fm_sum pre dátovú sadu rtu11



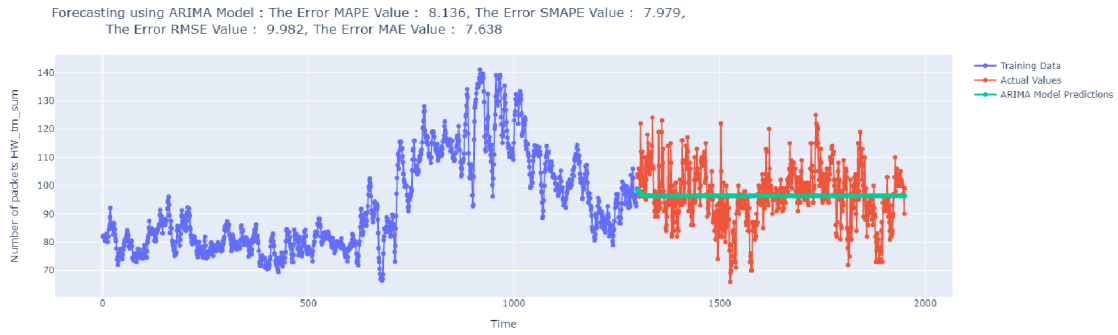
Obr. B.20: Graf predikcie modelu PROPHET časového radu fm_sum pre dátovú sadu rtu11



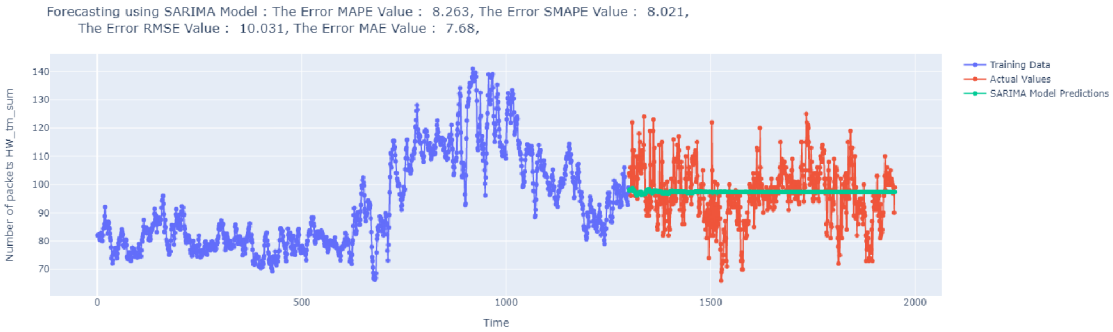
Obr. B.21: Graf predikcie modelu XGB časového radu fm_sum pre dátovú sadu rtu11



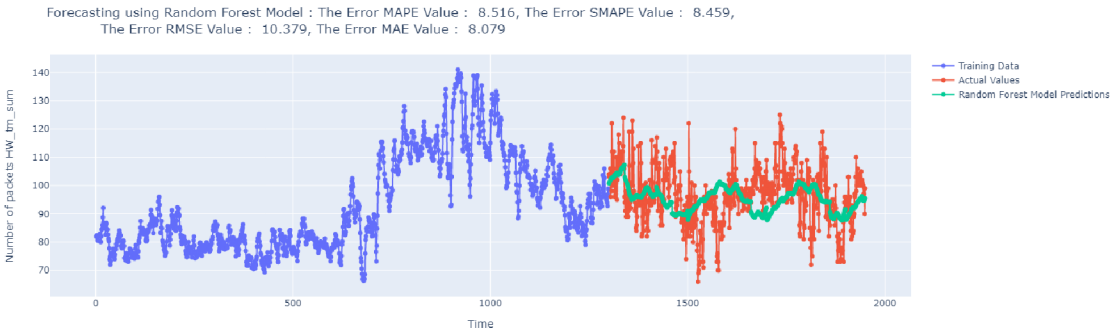
Obr. B.22: Graf predikcie modelu AR-X časového radu tm_sum pre dátovú sadu rtu11



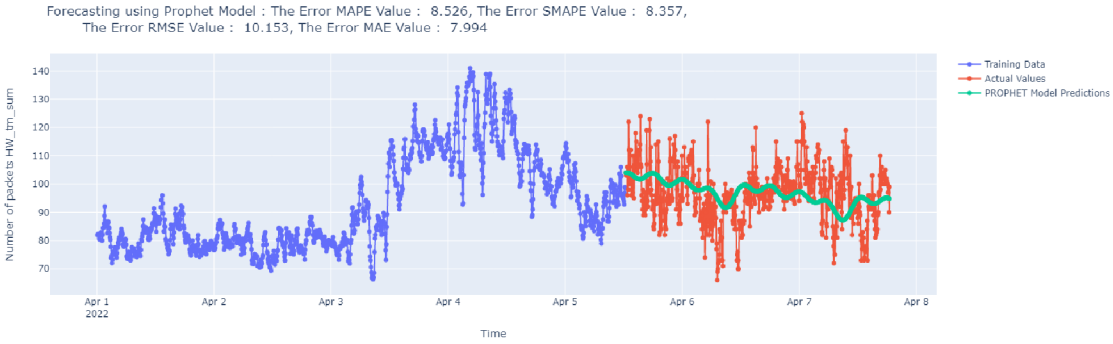
Obr. B.23: Graf predikcie modelu ARIMA časového radu tm_sum pre dátovú sadu rtu11



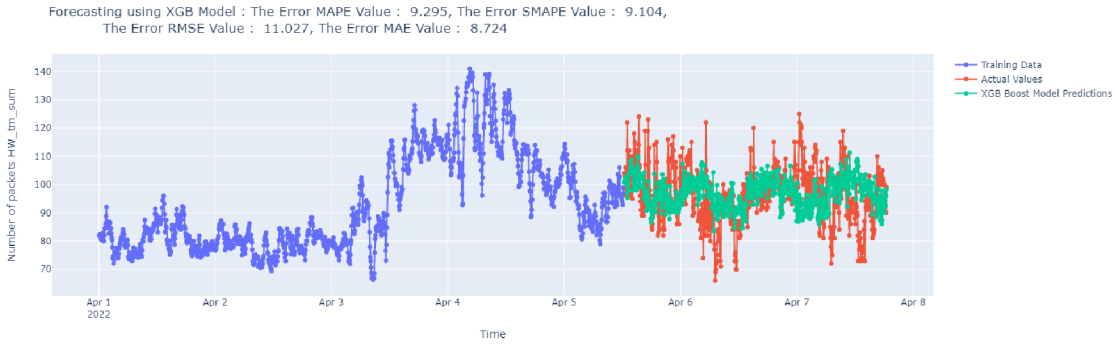
Obr. B.24: Graf predikcie modelu SARIMA časového radu tm_sum pre dátovú sadu rtu11



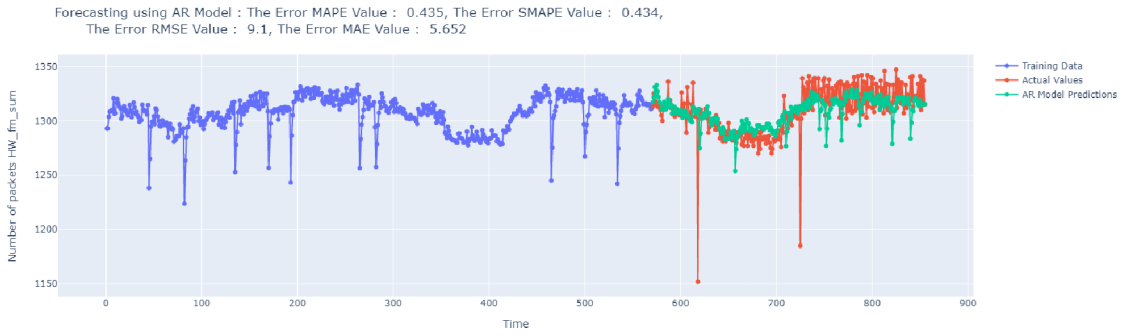
Obr. B.25: Graf predikcie modelu Random Forest časového radu tm_sum pre dátovú sadu rtu11



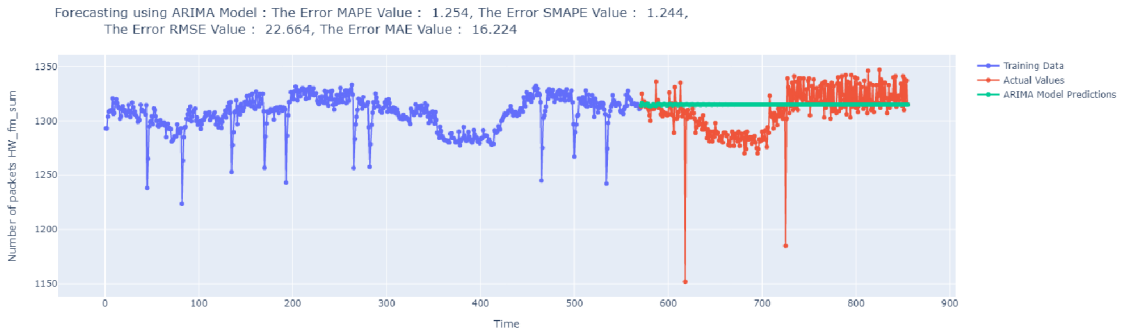
Obr. B.26: Graf predikcie modelu PROPHET časového radu tm_sum pre dátovú sadu rtu11



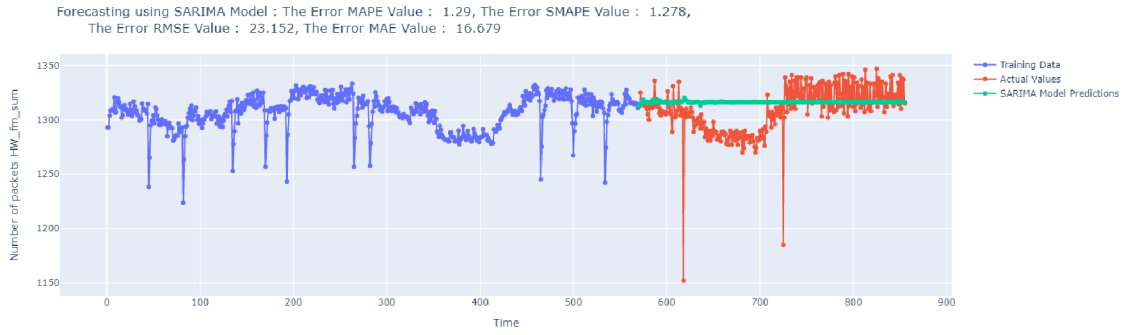
Obr. B.27: Graf predikcie modelu XGB časového radu tm_sum pre dátovú sadu rtu11



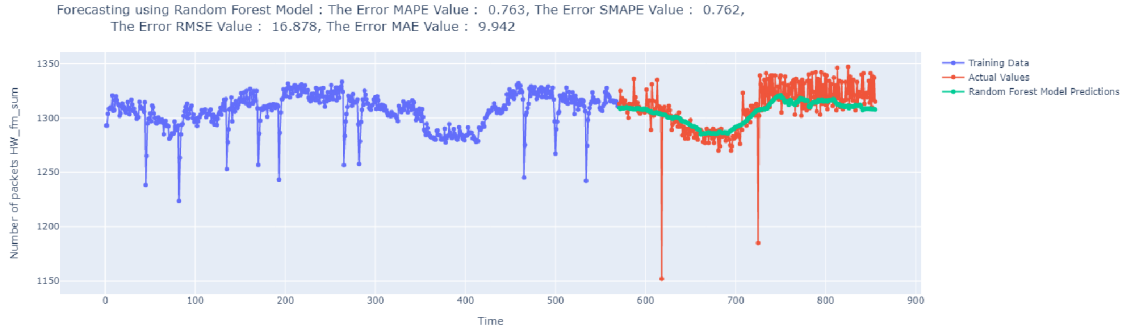
Obr. B.28: Graf predikcie modelu AR-X časového radu fm_sum pre dátovú sadu 104Mega



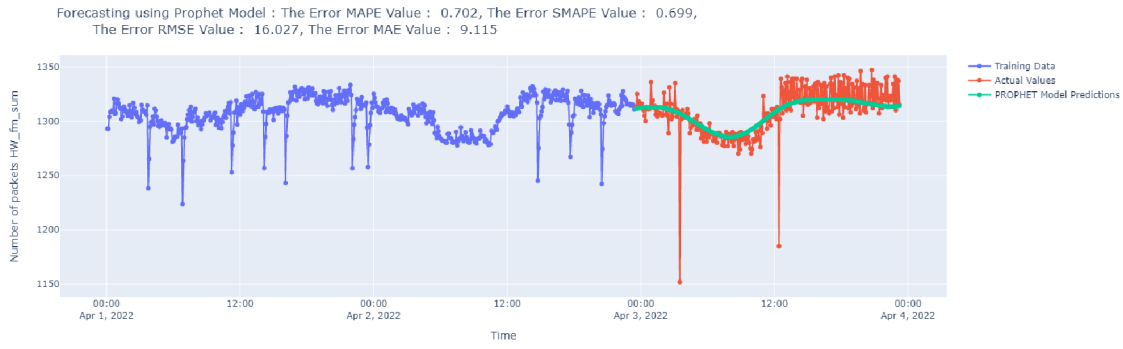
Obr. B.29: Graf predikcie modelu ARIMA časového radu fm_sum pre dátovú sadu 104Mega



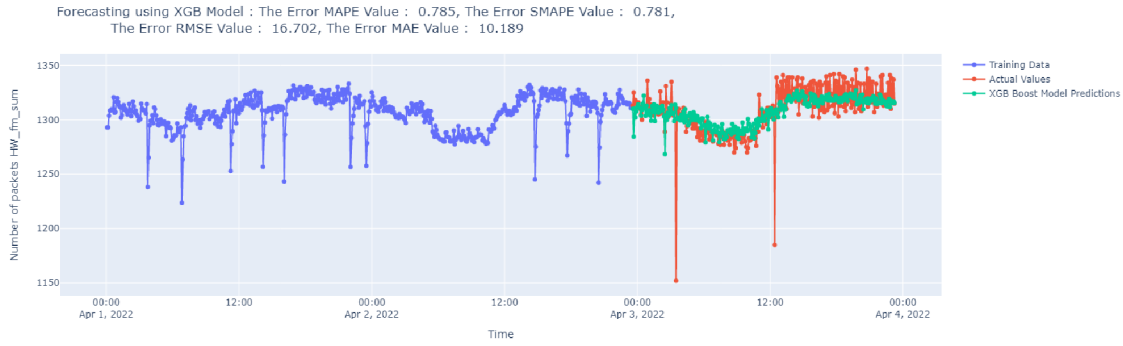
Obr. B.30: Graf predikcie modelu SARIMA časového radu fm_sum pre dátovú sadu 104Mega



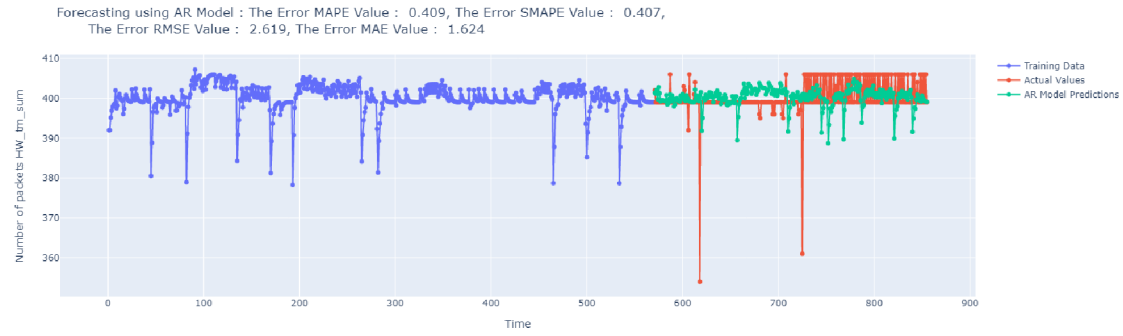
Obr. B.31: Graf predikcie modelu Random Forest časového radu fm_sum pre dátovú sadu 104Mega



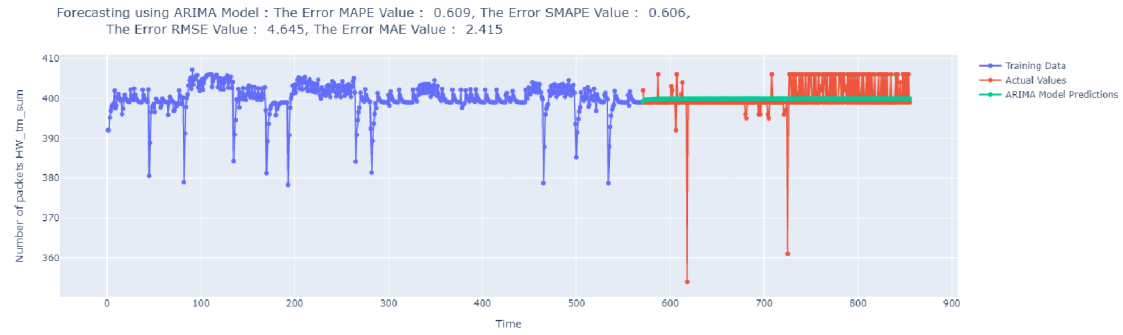
Obr. B.32: Graf predikcie modelu PROPHET časového radu fm_sum pre dátovú sadu 104Mega



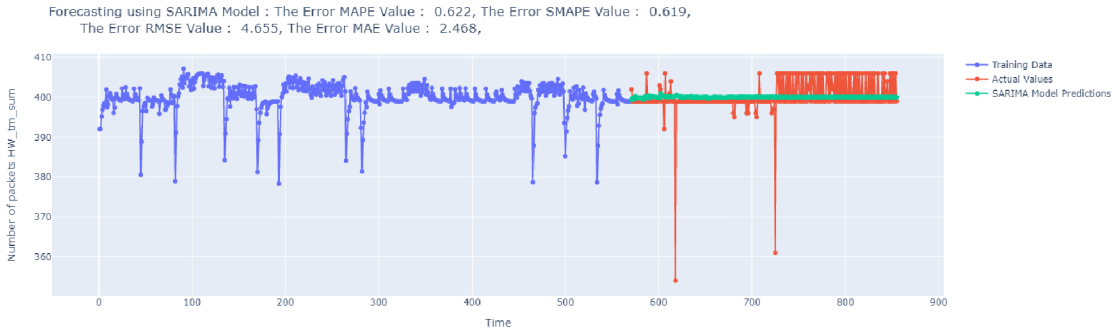
Obr. B.33: Graf predikcie modelu XGB časového radu fm_sum pre dátovú sadu 104Mega



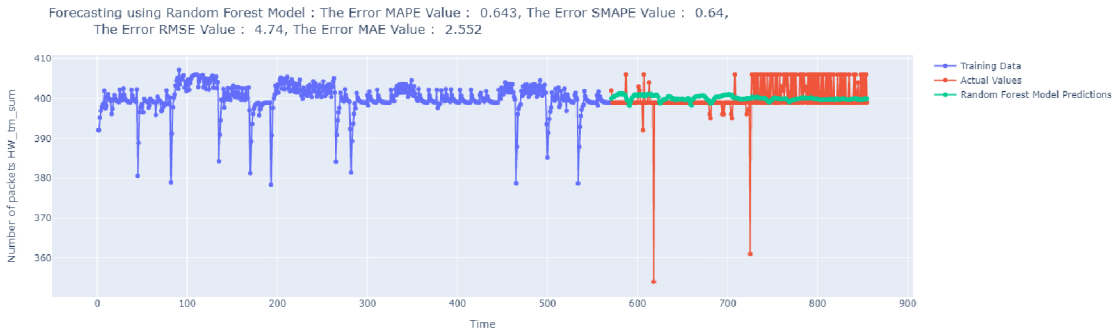
Obr. B.34: Graf predikcie modelu AR-X časového radu tm_sum pre dátovú sadu 104Mega



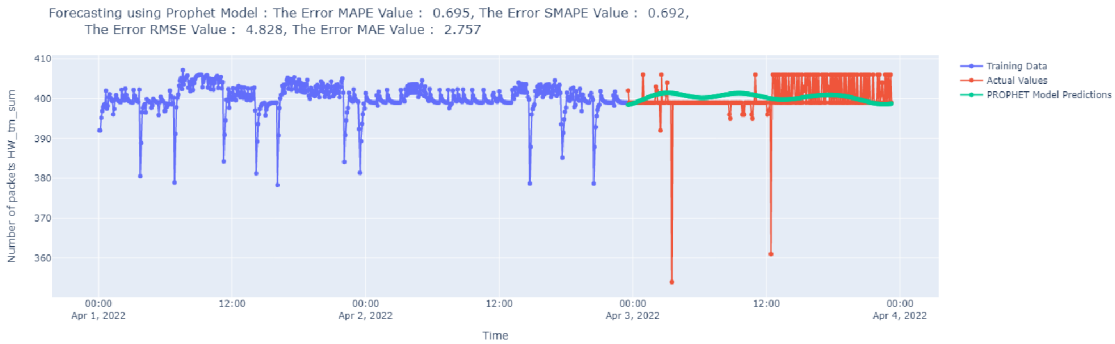
Obr. B.35: Graf predikcie modelu ARIMA časového radu tm_sum pre dátovú sadu 104Mega



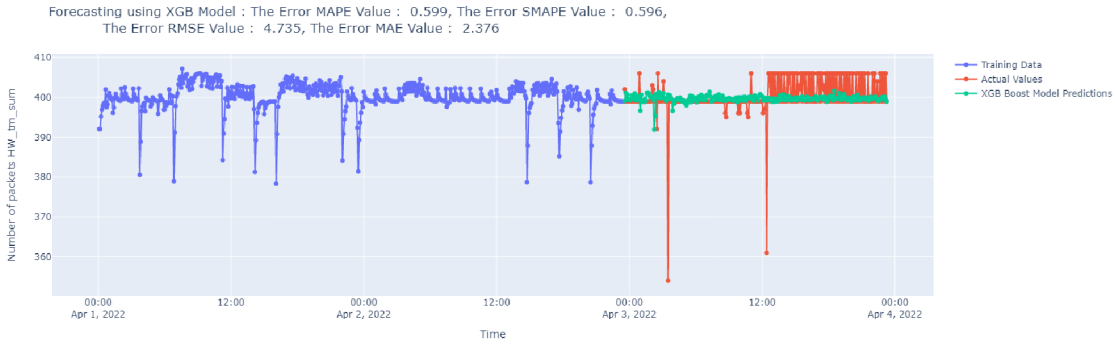
Obr. B.36: Graf predikcie modelu SARIMA časového radu tm_sum pre dátovú sadu 104Mega



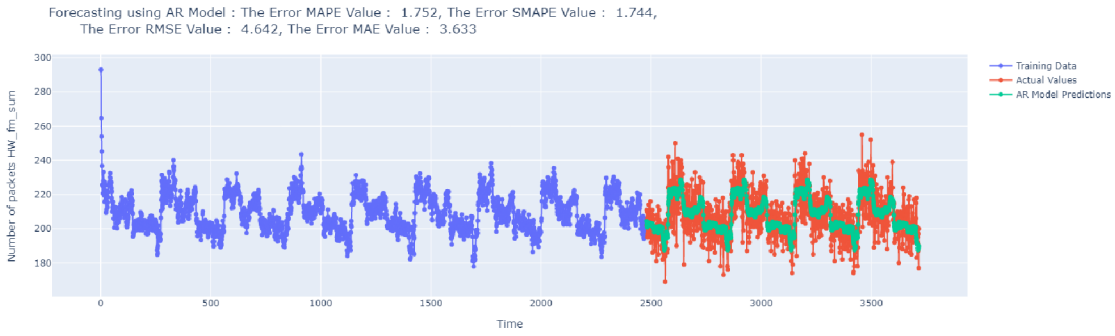
Obr. B.37: Graf predikcie modelu Random Forest časového radu tm_sum pre dátovú sadu 104Mega



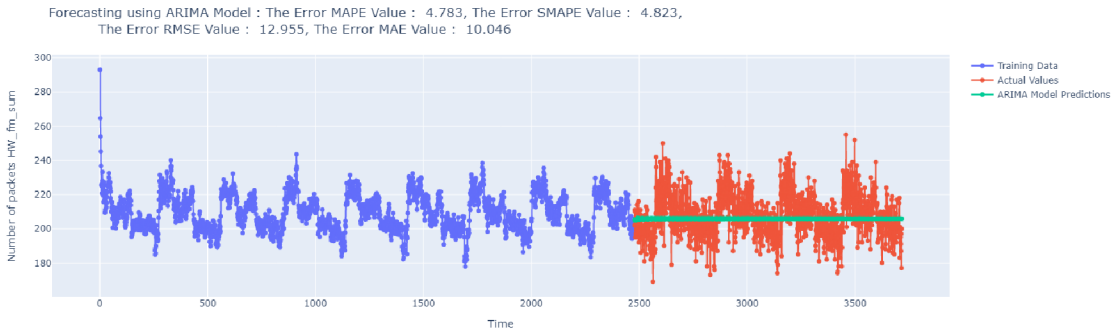
Obr. B.38: Graf predikcie modelu PROPHET časového radu tm_sum pre dátovú sadu 104Mega



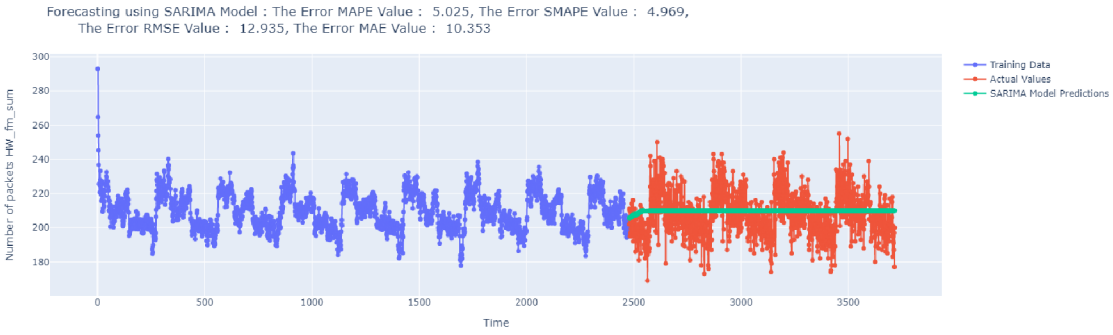
Obr. B.39: Graf predikcie modelu XGB časového radu tm_sum pre dátovú sadu 104Mega



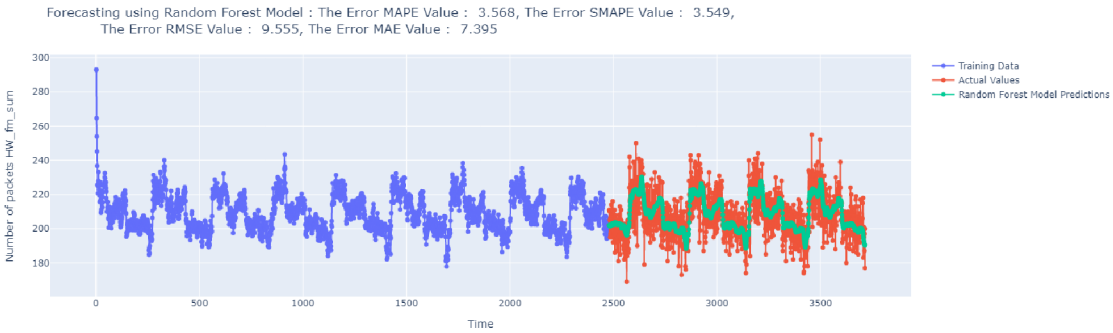
Obr. B.40: Graf predikcie modelu AR-X časového radu fm_sum pre dátovú sadu 10days



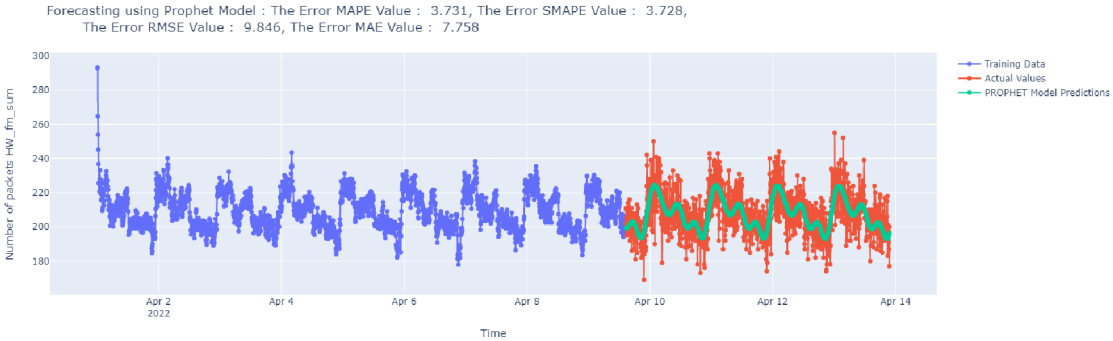
Obr. B.41: Graf predikcie modelu ARIMA časového radu fm_sum pre dátovú sadu 10days



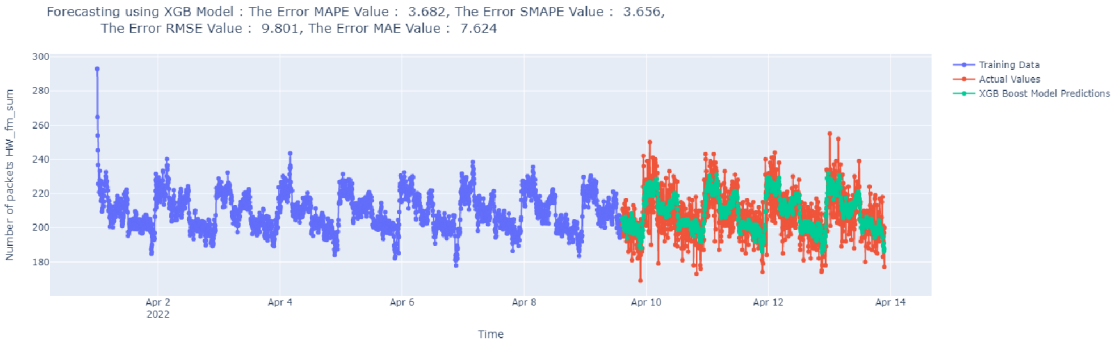
Obr. B.42: Graf predikcie modelu SARIMA časového radu fm_sum pre dátovú sadu 10days



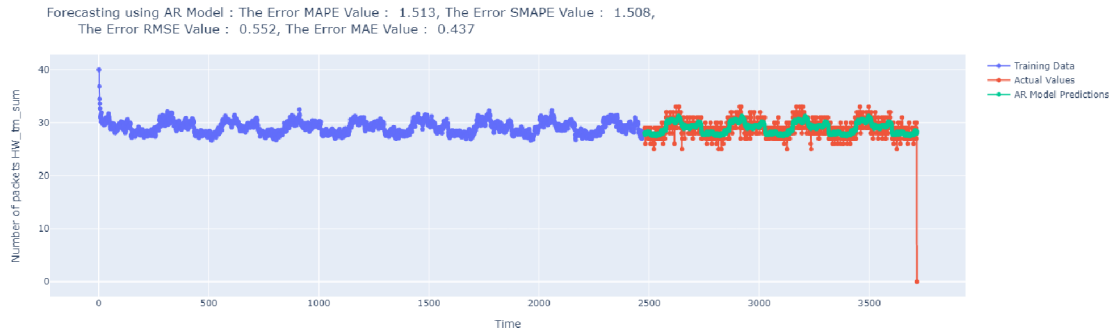
Obr. B.43: Graf predikcie modelu Random Forest časového radu fm_sum pre dátovú sadu 10days



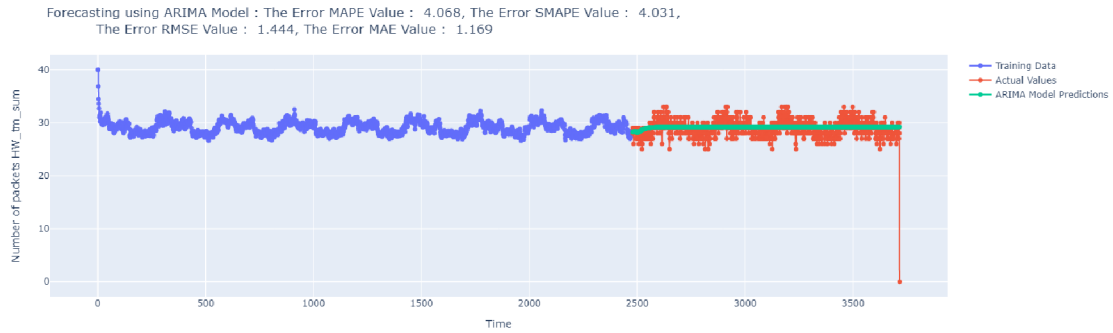
Obr. B.44: Graf predikcie modelu PROPHET časového radu fm_sum pre dátovú sadu 10days



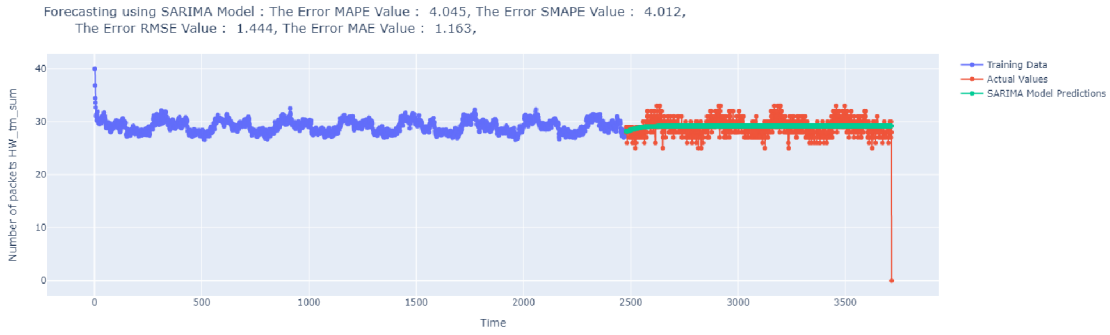
Obr. B.45: Graf predikcie modelu XGB časového radu fm_sum pre dátovú sadu 10days



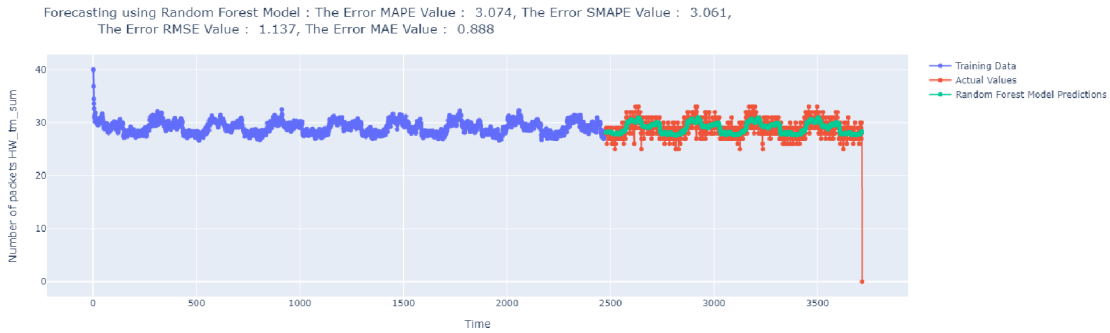
Obr. B.46: Graf predikcie modelu AR-X časového radu tm_sum pre dátovú sadu 10days



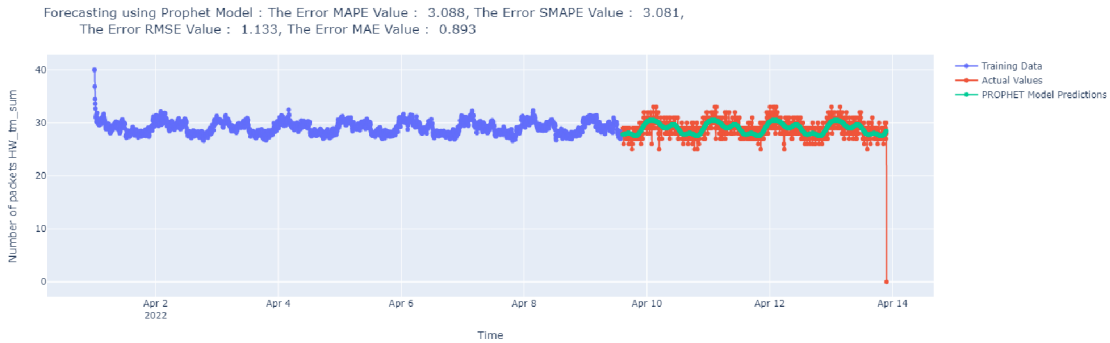
Obr. B.47: Graf predikcie modelu ARIMA časového radu tm_sum pre dátovú sadu 10days



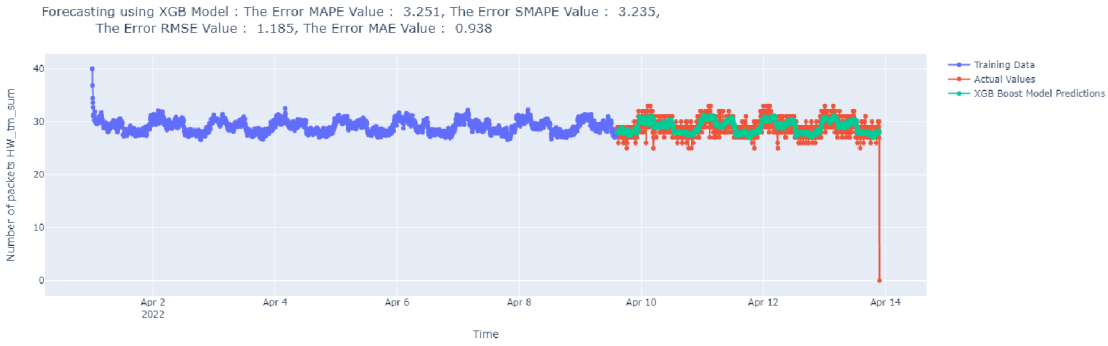
Obr. B.48: Graf predikcie modelu SARIMA časového radu tm_sum pre dátovú sadu 10days



Obr. B.49: Graf predikcie modelu Random Forest časového radu tm_sum pre dátovú sadu 10days



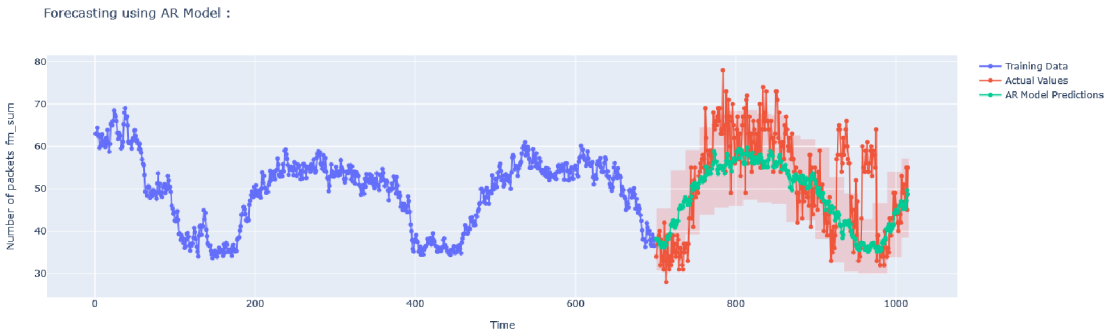
Obr. B.50: Graf predikcie modelu PROPHET časového radu tm_sum pre dátovú sadu 10days



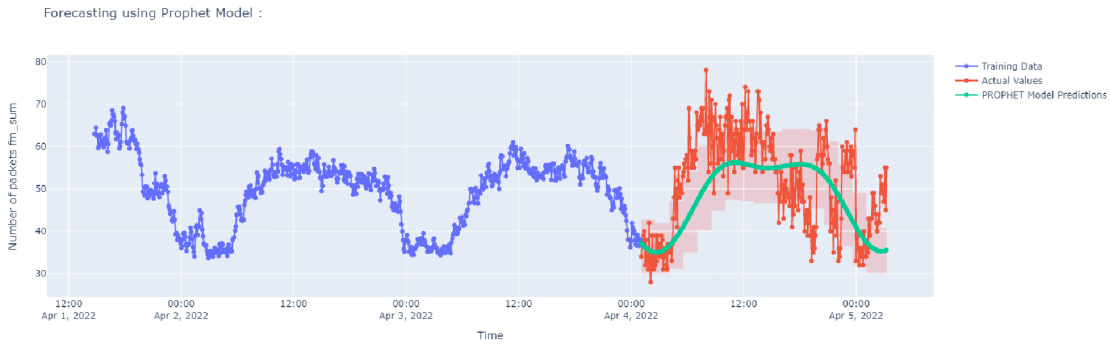
Obr. B.51: Graf predikcie modelu XGB časového radu tm_sum pre dátovú sadu 10days

Príloha C

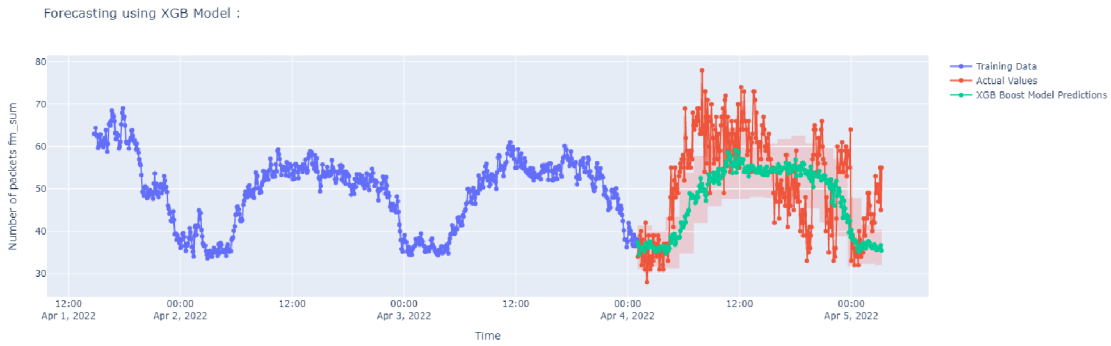
Vizualizácia detekcie útokov



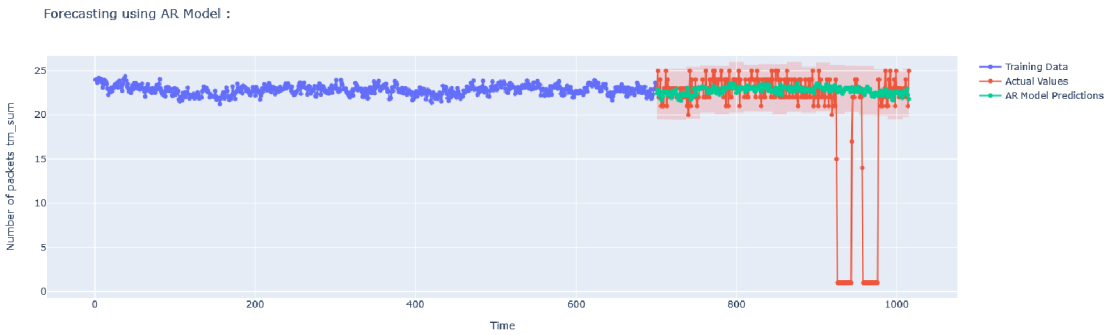
Obr. C.1: Graf detekcie dos útoku za použitia modelu AR-X v časovom rade fm_sum



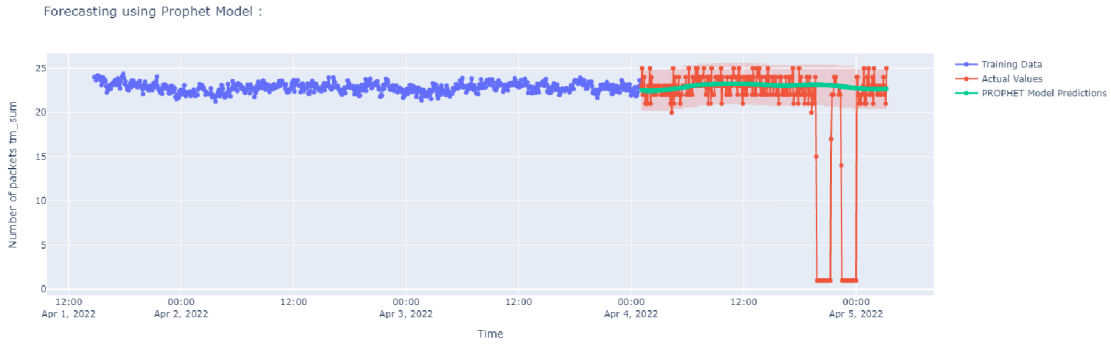
Obr. C.2: Graf detekcie dos útoku za použitia modelu PROPHET v časovom rade fm_sum



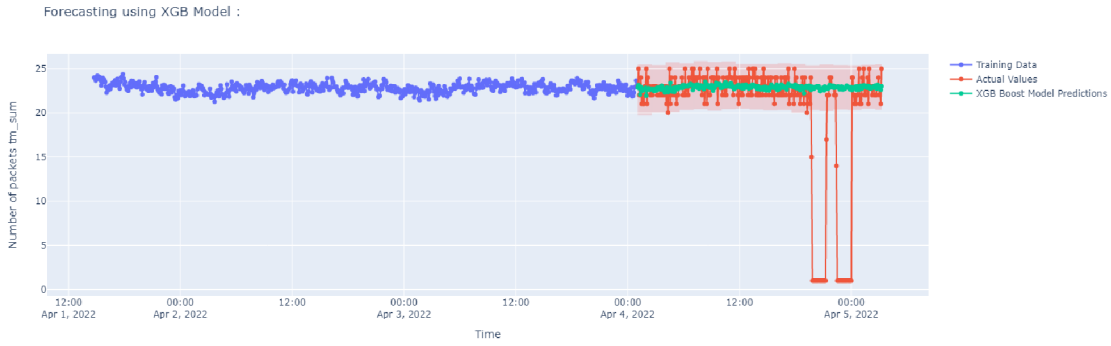
Obr. C.3: Graf detekcie dos útoku za použitia modelu XGB v časovom rade fm_sum



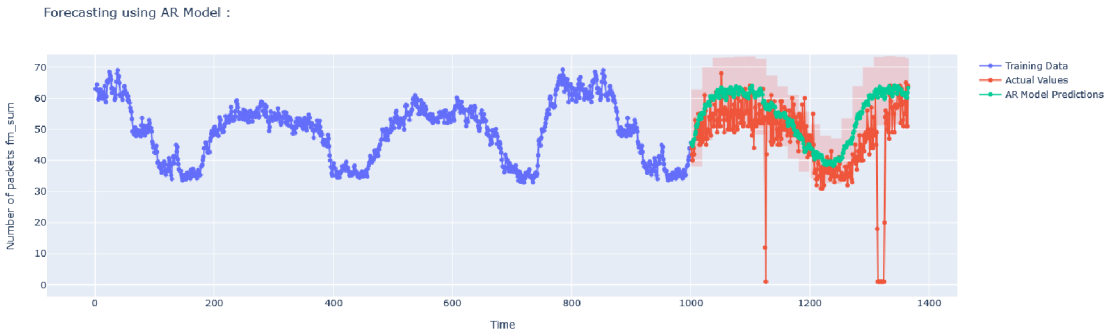
Obr. C.4: Graf detekcie dos útoku za použitia modelu AR-X v časovom rade tm_sum



Obr. C.5: Graf detekcie dos útoku za použitia modelu PROPHET v časovom rade tm_sum

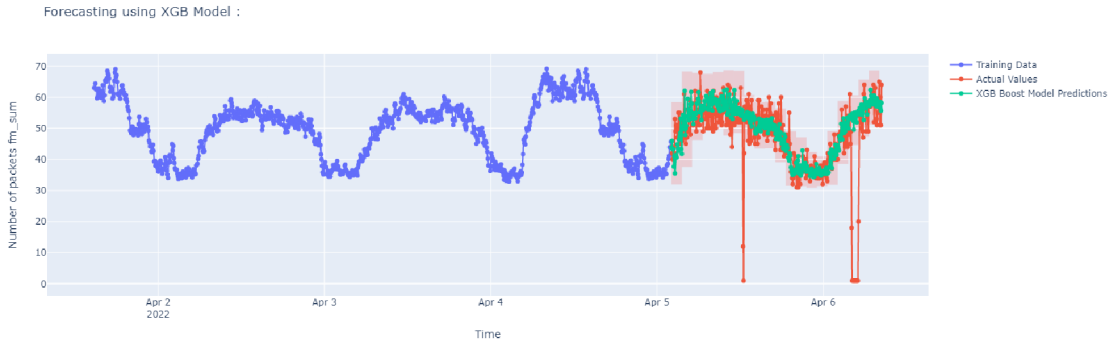


Obr. C.6: Graf detekcie dos útoku za použitia modelu XGB v časovom rade tm_sum



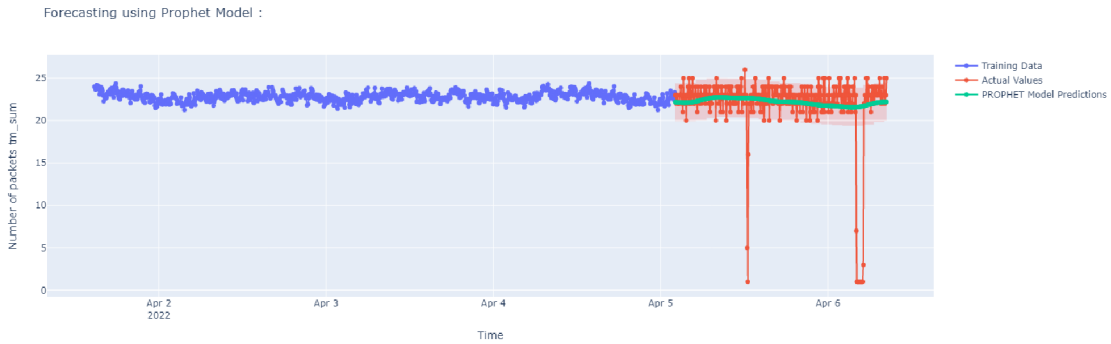
Obr. C.7: Graf detekcie straty spojenia za použitia modelu AR-X v časovom rade fm_sum

Obr. C.8: Graf detekcie straty spojenia za použitia modelu PROPHET v časovom rade fm_sum

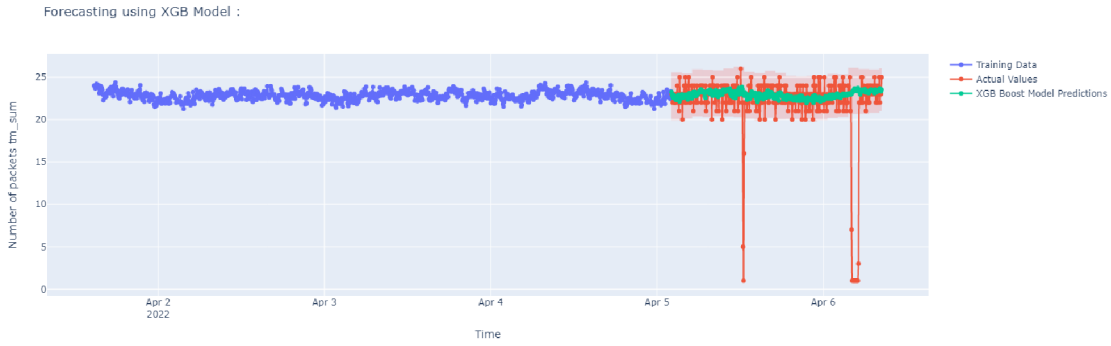


Obr. C.9: Graf detekcie straty spojenia za použitia modelu XGB v časovom rade fm_sum

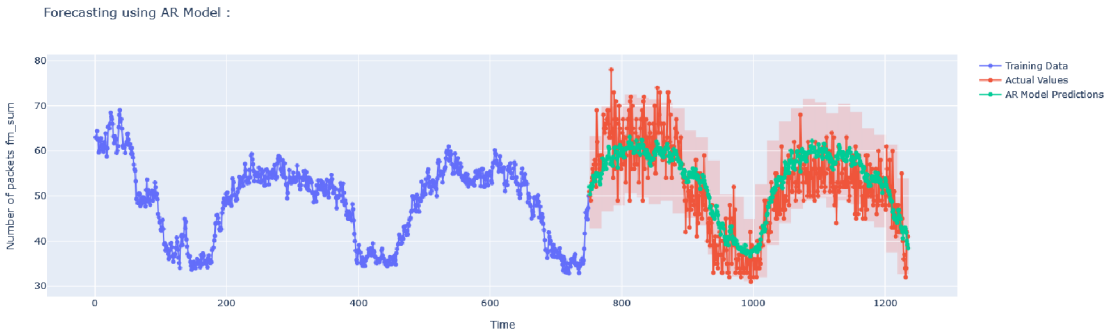
Obr. C.10: Graf detekcie straty spojenia za použitia modelu AR-X v časovom rade tm_sum



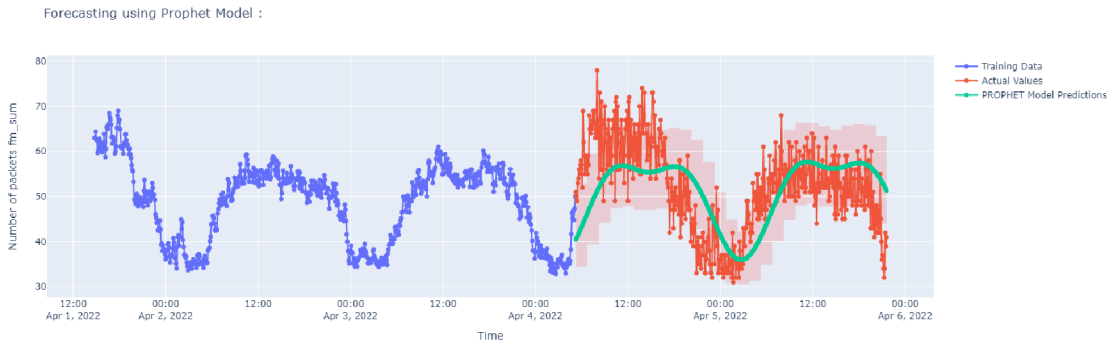
Obr. C.11: Graf detekcie straty spojenia za použitia modelu PROPHET v časovom rade tm_sum



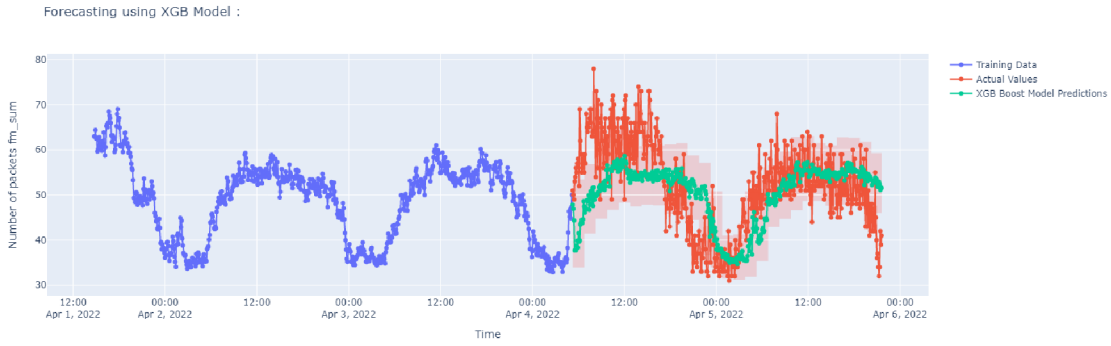
Obr. C.12: Graf detekcie straty spojenia za použitia modelu XGB v časovom rade tm_sum



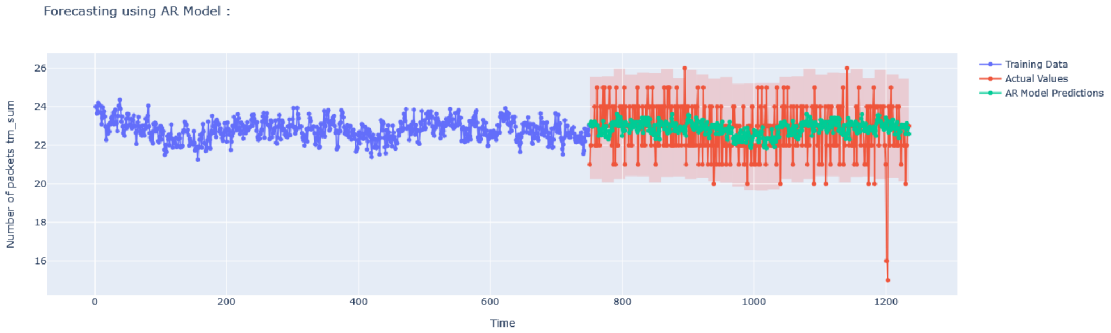
Obr. C.13: Graf detekcie útoku pomocou injekcie za použitia modelu AR-X v časovom rade fm_sum



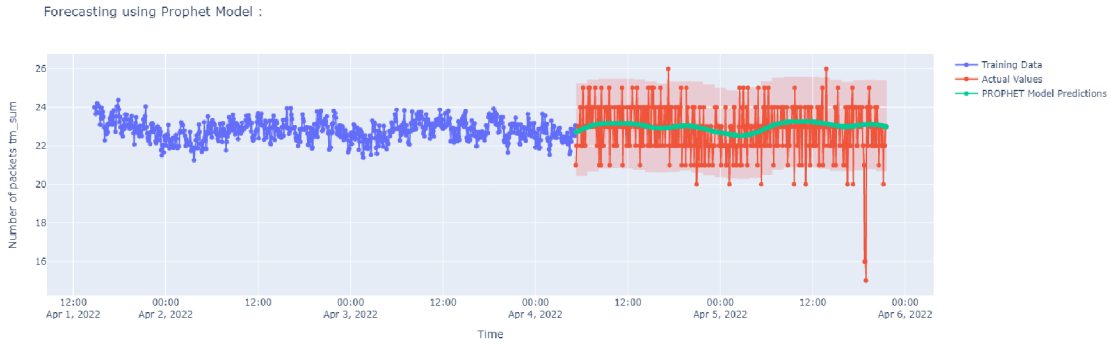
Obr. C.14: Graf detekcie útoku pomocou injekcie za použitia modelu PROPHET v časovom rade fm_sum



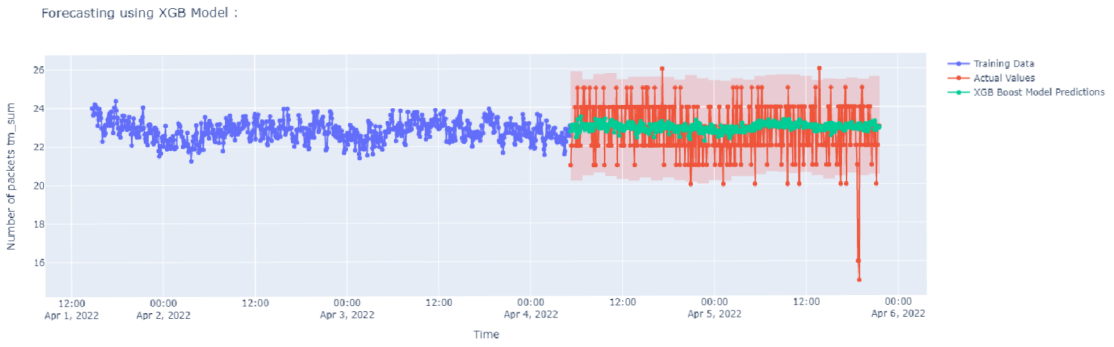
Obr. C.15: Graf detekcie útoku pomocou injekcie za použitia modelu XGB v časovom rade fm_sum



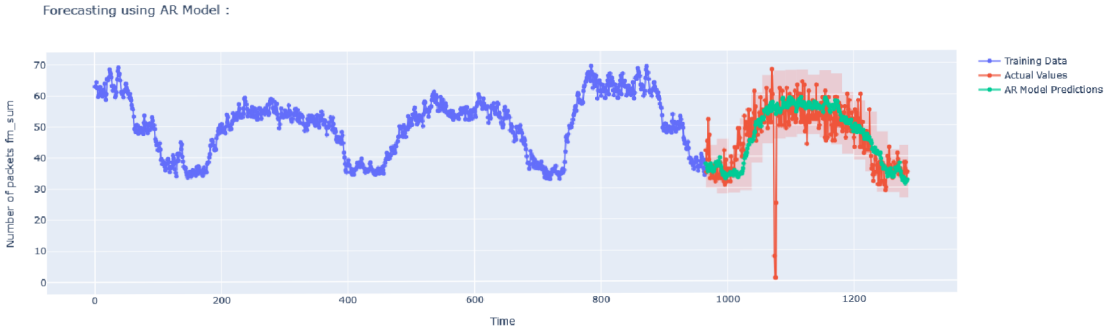
Obr. C.16: Graf detekcie útoku pomocou injekcie za použitia modelu AR-X v časovom rade tm_sum



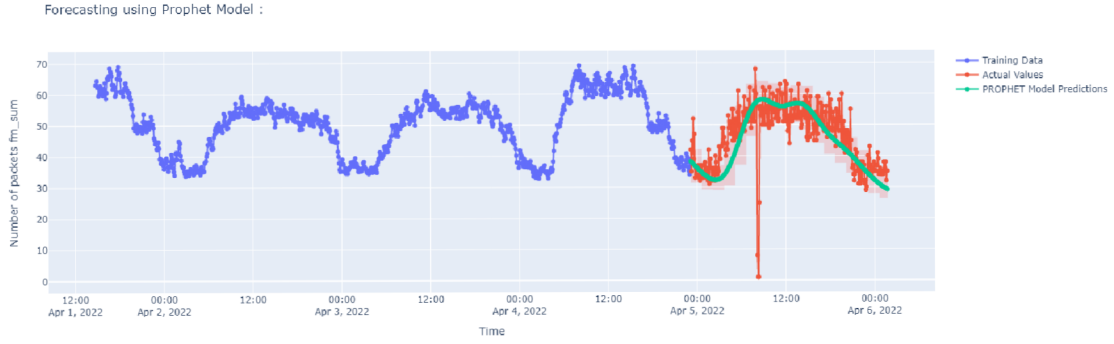
Obr. C.17: Graf detekcie útoku pomocou injekcie za použitia modelu PROPHET v časovom rade tm_sum



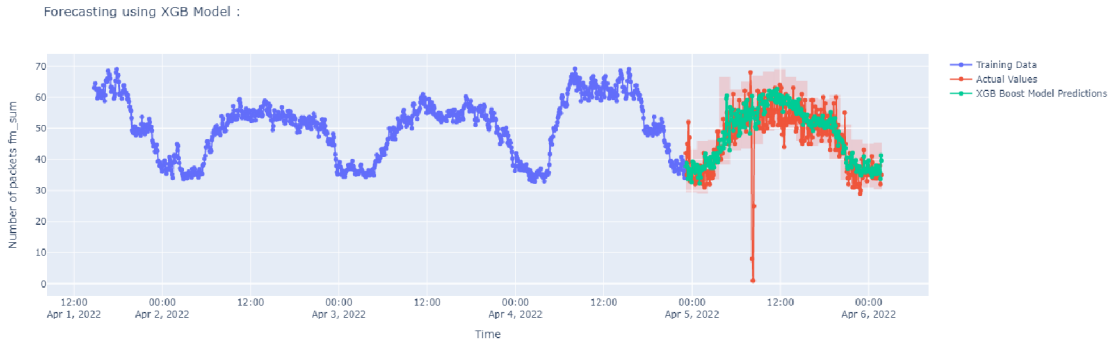
Obr. C.18: Graf detekcie útoku pomocou injekcie za použitia modelu XGB v časovom rade tm_sum



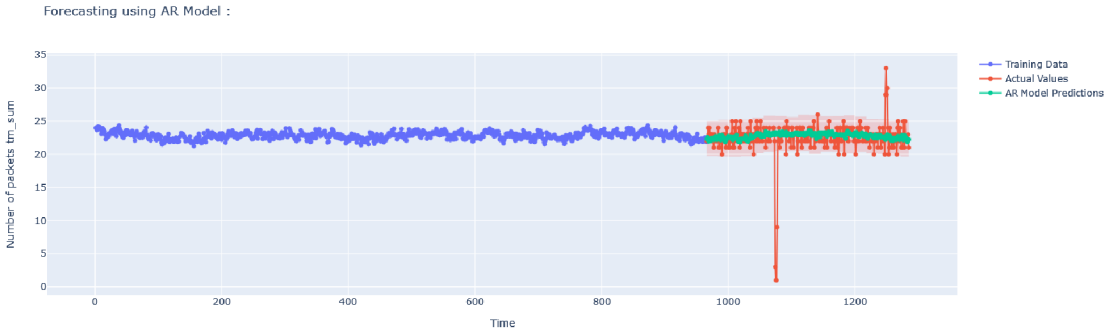
Obr. C.19: Graf detekcie útoku skenovaním za použitia modelu AR-X v časovom rade fm_sum



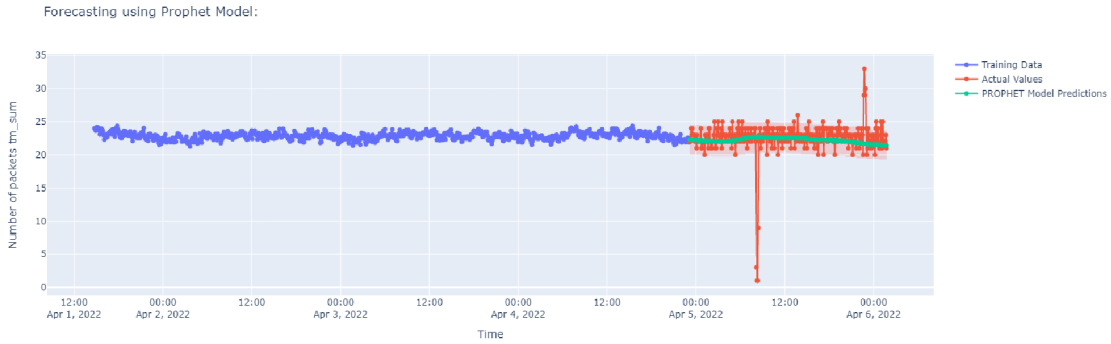
Obr. C.20: Graf detekcie útoku skenovaním za použitia modelu PROPHET v časovom rade fm_sum



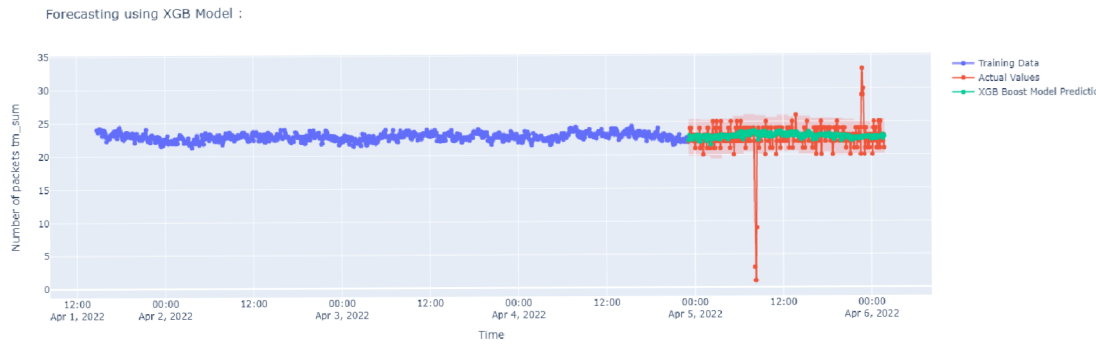
Obr. C.21: Graf detekcie útoku skenovaním za použitia modelu XGB v časovom rade fm_sum



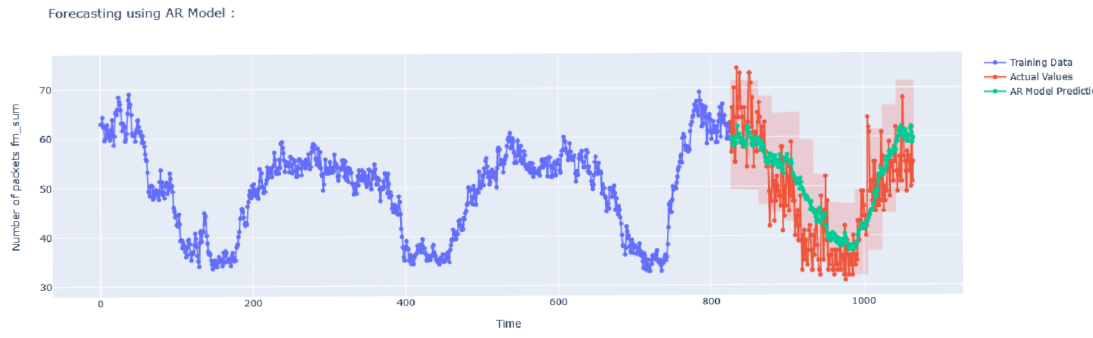
Obr. C.22: Graf detekcie útoku skenovaním za použitia modelu AR-X v časovom rade tm_sum



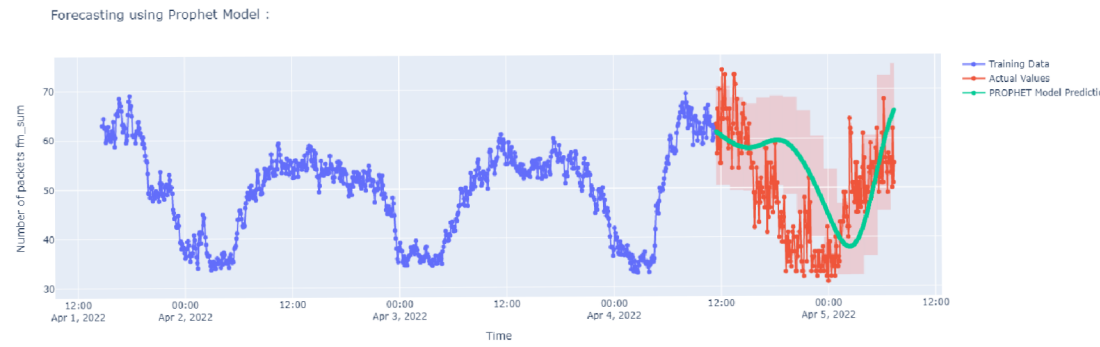
Obr. C.23: Graf detekcie útoku skenovaním za použitia modelu PROPHET v časovom rade tm_sum



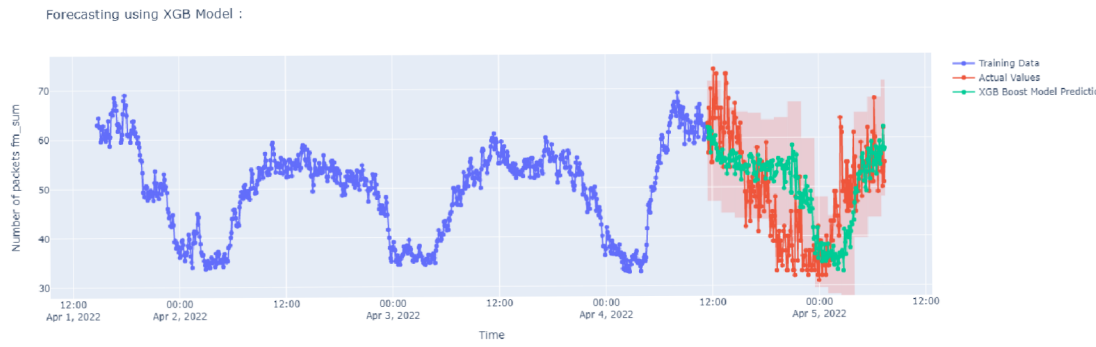
Obr. C.24: Graf detekcie útoku skenovaním za použitia modelu XGB v časovom rade tm_sum



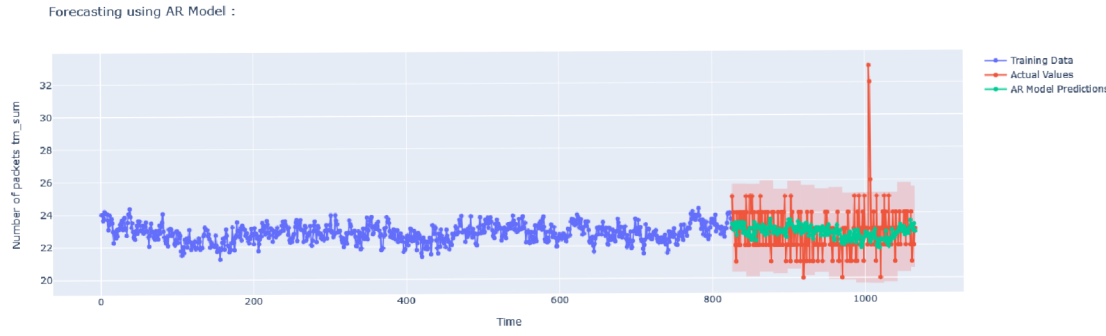
Obr. C.25: Graf detekcie útoku vypínaním a zapínaním zariadenia za použitia modelu AR-X v časovom rade fm_sum



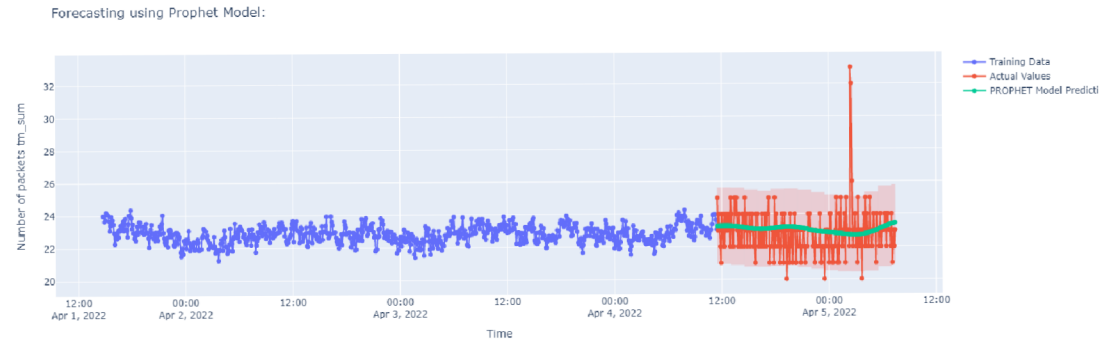
Obr. C.26: Graf detekcie útoku vypínaním a zapínaním zariadenia za použitia modelu PROPHET v časovom rade fm_sum



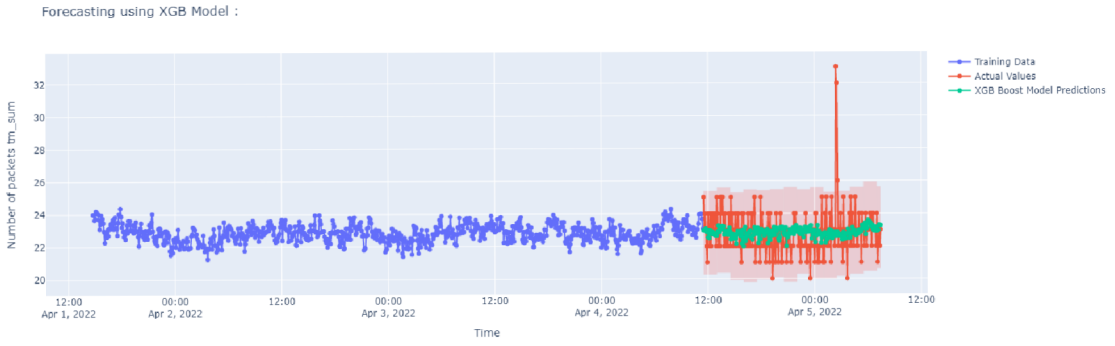
Obr. C.27: Graf detekcie útoku vypínaním a zapínaním zariadenia za použitia modelu XGB v časovom rade fm_sum



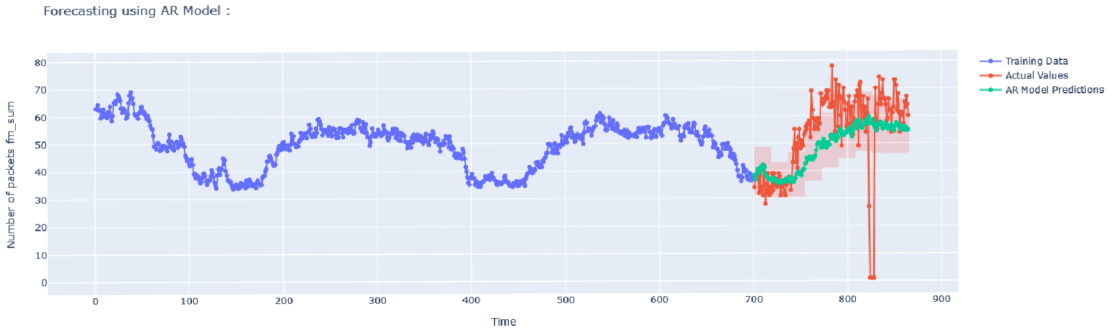
Obr. C.28: Graf detekcie útoku vypínaním a zapínaním zariadenia za použitia modelu AR-X v časovom rade tm_sum



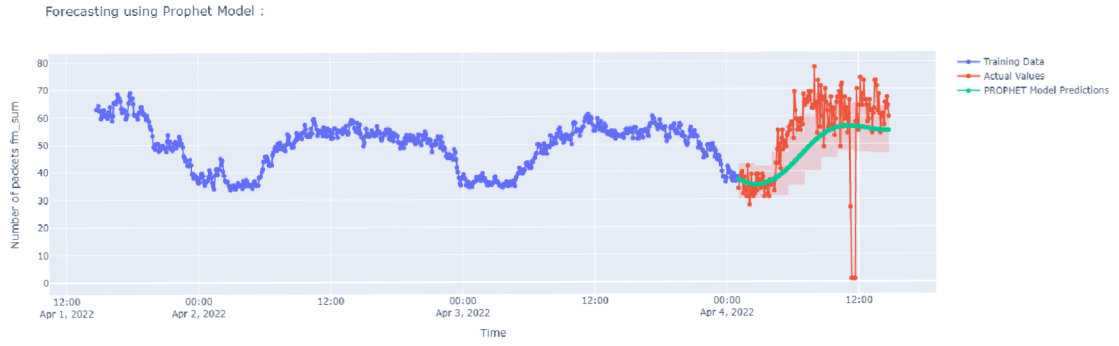
Obr. C.29: Graf detekcie útoku vypínaním a zapínaním zariadenia za použitia modelu PROPHET v časovom rade tm_sum



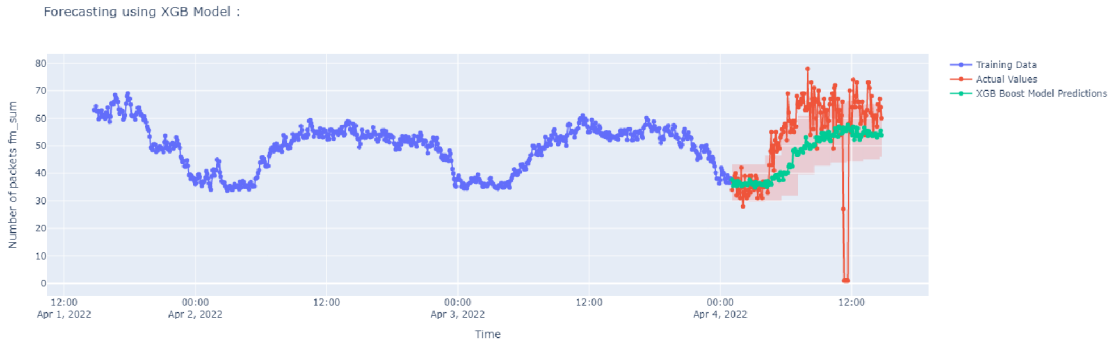
Obr. C.30: Graf detekcie útoku vypínaním a zapínaním zariadenia za použitia modelu XGB v časovom rade tm_sum



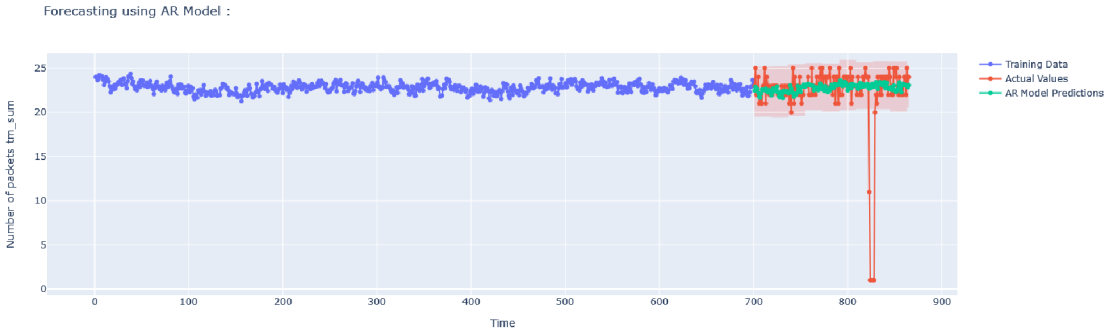
Obr. C.31: Graf detekcie útoku rogue device za použitia modelu AR-X v časovom rade fm_sum



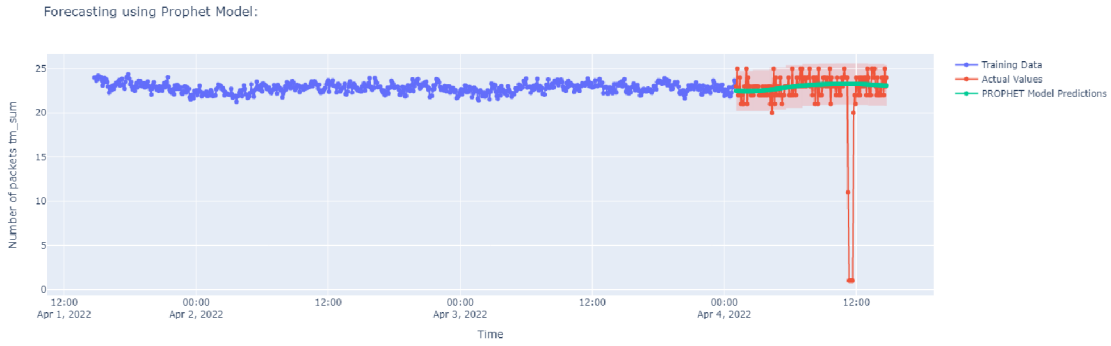
Obr. C.32: Graf detekcie útoku rogue device za použitia modelu PROPHET v časovom rade fm_sum



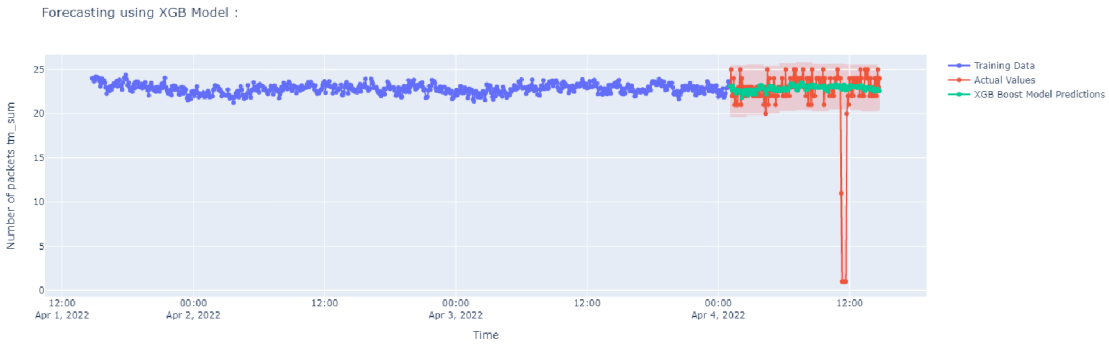
Obr. C.33: Graf detekcie útoku rogue device za použitia modelu XGB v časovom rade fm_sum



Obr. C.34: Graf detekcie útoku rogue device za použitia modelu AR-X v časovom rade tm_sum



Obr. C.35: Graf detekcie útoku rogue device za použitia modelu PROPHET v časovom rade tm_sum



Obr. C.36: Graf detekcie útoku rogue device za použitia modelu XGB v časovom rade tm_sum

Príloha D

Obsah DVD

Priložené DVD obsahuje:

- latex/ - zdrojové súbory tejto práce pre LaTeX
- experiments/ - Zdrojové súbory v podobe Jupyter Notebook-ov, použité pre experimentálnu časť tejto práce, v tomto priečinku sa nachádza aj priečinok s datasetmi.
- experiments/data - datasety použité vrámci tejto práce

Výstup experimentov je možné preskúmať v prílohách B a C alebo získať spustením jednotlivých Notebook-ov. V tomto prípade sa vytvorí zložka Visualization, do ktorej sa uložia príslušné grafy v podobe html súborov. Html formát je zvolený z dôvodu, aby sa zachovala možnosť interagovať s jednotlivými grafmi. Pamäťové nároky tohto priečinka dosahovali takmer 400MB, preto nie sú predom vygenerované v tejto zložke.