KAUNO TECHNOLOGIJOS UNIVERSITETAS

MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS

Rokas Vasiliūnas

Anomalijų detekcija e-prekyboje: "Little Data" atvejis

Baigiamasis magistro projektas

Vadovai
Doc. dr. Audrius Kabašinskas
Prof. dr. Rimantas Gatautis

KAUNAS, 2017

KAUNO TECHNOLOGIJOS UNIVERSITETAS

MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS

Anomalijų detekcija e-prekyboje: „Little Data“ atvejis

Baigiamasis magistro projektas

Didžiųjų verslo duomenų analitika (kodas 621G12002)

Vadovai
(parašas) Doc. dr. Audrius Kabašinskas
(data)
(parašas) Prof. dr. Rimantas Gatautis
(data)
Recenzentai
(parašas) Doc. dr. Aistė Dovalienė
(data)
(parašas) Doc. dr. Loreta Saunorienė
(data)

Projektą atliko
(parašas) Rokas Vasiliūnas
(data)

KAUNAS, 2017

KAUNO TECHNOLOGIJOS UNIVERSITETAS

Matematikos ir gamtos mokslų fakltetas

(Fakultetas)

Rokas Vasiliūnas

(Studento vardas, pavardė)

Didžiųjų verslo duomenų analitika, 621G12002

(Studijų programos pavadinimas, kodas)

„Baigiamojo projekto pavadinimas“

AKADEMINIO SĄŽININGUMO DEKLARACIJA

20 ___ m. _____ ____ d.

Kaunas

Patvirtinu, kad mano, **Roko Vasiliūno**, baigiamasis projektas tema „Anomalijų detekcija e-prekyboje: „Little Data“ atvejis“ yra parašytas visiškai savarankiškai ir visi pateikti duomenys ar tyrimų rezultatai yra teisingi ir gauti sąžiningai. Šiame darbe nei viena dalis nėra plagijuota nuo jokių spausdintinių ar internetinių šaltinių, visos kitų šaltinių tiesioginės ir netiesioginės citatos nurodytos literatūros nuorodose. Įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs.

Aš suprantu, kad išaiškėjus nesąžiningumo faktui, man bus taikomos nuobaudos, remiantis Kauno technologijos universitete galiojančia tvarka.

_____ _____

(vardą ir pavardę įrašyti ranka) *(parašas)*

TURINYS

Vasiliūnas, Rokas. „Anomalijų detekcija e-prekyboje: „Little Data“ atvejis. *Magistro* baigiamasis projektas / vadovai doc. dr. Audrius Kabašinskas ir Prof. Dr. Rimantas Gatautis; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

Fiziniai mokslai, Matematika (01 P)

Reikšminiai žodžiai: *anomalija, elektroninė prekyba, laiko eilutės, arima, var.*

Kaunas, 2017. 82 p.

SANTRAUKA

Šiame darbe yra atliktas anomalijų tyrimas “Little Data” duomenų rinkiniui. Internetinė prekyba yra labai konkurencinga, todėl yra svarbu žinoti veiksnius, kurie lemia pirkėjo pasirinkimą. Anomalijų aptikimas gali padėti rasti pirkėjų pasirinkimą lemiančius veiksnius ir padėti juos panaudoti pardavėjo naudai. Duomenis sudaro 9 kintamieji iš kurių 7 buvo naudojami tyrimams. Duomenys yra dieniniai – 732 eilutės, iš kurių 26 dienos yra priskiriamos anomalijoms. Nagrinėjamas duomenų laikotarpis – 2014-06-30 – 2016-07-01. Duomenims yra būdingas nestacionarumas ir sezoniškumas, todėl duomenys buvo diferencijuojami ir pritaikyti sezoniškumui.

Tyrimas buvo atliekamas 3 metodais – autoregresijos su slenkančiu vidurkiu (ARIMA), vektorinės autoregresijos (VAR) ir naudojantis “AnomalyDetection” R paketą. Kiekvienas modelis buvo sudarytas panaudojant tris skirtingus pasikliautinumo intervalus anomalijoms aptikti: 90 %, 95 % ir 99 %. Kiekvieno modelio tyrimas buvo atliekamas padalinus kintamuosius į atskiras laiko eilutes po vienerius metus. Pabaigoje buvo sudedamos visos rastos anomalijos. Duomenys vertinami remiantis sumaišymo matrica ir dvejais įvertinimais – specifiškumu ir tikslumu.

ARIMA metodas aptiko 14 (99 % pasikliautinasis intervalas), 16 (95 % pasikliautinasis intervalas) ir 18 (90 % pasikliautinasis intervalas) anomalijų, VAR su atitinkaais intervalais – 15, 17 ir 20 anomalijų, o “AnomalyDetection” tik 13, 14 ir 14 anomalijų. Pagal gautus rezultatus VAR metodas yra geriausias. Po tyrimo buvo padaryta išvada, kad norint tikslesnių rezultatų reikia daugiau duomenų apie anomalijas. Taip pat tyrimui padėtų ilgesnio laikotarpio laiko eilutės.

Vasiliūnas, Rokas. Anomalies Detection in E-commerce: Case of "Little Data": *Master's* thesis in Applied Mathematics / supervisor assoc. prof. Audrius Kabašinskas and Prof.dr. Rimantas Gatautis. The Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

Natural Sciences, Mathematics (01 P)

Key words: anomaly, detection, e-commerce, time, series;

Kaunas, 2017. 82 p.

SUMMARY

This thesis main goal is to detect anomalies in "Little Data" data set. E-commerce is very competitive, because of that, it is very important to know what determines customer decision to buy or not. Anomaly detection may help to find new determinants. Anomalies are big, sudden unknown changes in data. Ability to know causes of anomaly, gives the ability to control it.

Data set contains 9 variables, but only 7 of them are used in analysis. It's two years daily data from 2014-06-30 to 2016-07-01. In 732 rows of data, there is 26 anomalies. Default data is unstationary and has seasonal affect. Because of that data set was differenciated and seasonal adjusted.

There was used three models in anomaly detection – Auto Regressive Integrated Moving Average with Exogeneous Input (ARIMA), Vector autoregression with Exogeneous Input (VAR) models and R package called "AnomalyDetection". With every method there was used three dfferent confidence intervals (90 %, 95 %, 99 %). In every analysis data was divided by year and variable. At the end, results from all variables was summed. To evaluate models confussion matrix and its metrics (specificity and precision) were used.

ARIMA model detected 14 anomalies with 99 % confidence intervals, 16 with 95 % confidence intervals and 18 with 90 % confidence intervals. VAR model detected 15 anomalies with 99 % confidence intervals, 17 with 95 % confidence intervals and 20 with 90 % confidence intervals. R package "AnomalyDetection" got the worst result – 13 anomalies with 99 % confidence intervals, 14 with 95 % confidence intervals and 14 with 90 % confidence intervals. To sum up, we can say that all models were average in anomaly detection. As a recommendation, longer time series and more data about anomalies would help to improve model accuracy.

SANTRUMPOS

ARIMA – autoregresija su slenkančiu vidurkiu;

VAR – vektorinė autoregresija;

ĮŽANGA

Temos aktualumas: Elektroninė prekyba - prekybos būdas, kai prekių ar paslaugų pirkimas ir pardavimas yra vykdomas naudojantis elektroninėmis priemonėmis. Elektroninė prekyba yra plėtojama ne vieną dešimtį metų, tačiau tik 2002 metai yra laikomi tikrosios el. komersijos pradžia. 2015 metais trečdalis (32 %) Lietuvos Respublikos gyventojų pirko prekes arba paslaugas internetu. Tuo tarpu visoje Europos Sąjungoje tais pačiais metais 65 % gyventojų buvo pasinaudoję el. prekybos paslaugomis. Per 2016 metus žmonių skaičius dar labiau padidėjo. Elektronikos prietaisų populiarumas ir kasdieninis naudojimasis iššaukė, kad praeitas metais (2016), JAV buvo daugiau pirkėjų, kurie užsakymus įvykdė per mobiliuosius telefonus negu per asmeninius kompiuterius. Spartus el. prekybos augimas skatina bendroves tyrinėti internetinių pirkėjų elgesį ir jo priežastis. Anomalijų aptikimas – gali būti vienas iš būdų geriau pažinti pirkėjus. Anomalija el. prekyboje yra ryškus statistikos rodiklių pasikeitimai dėl nežinomų priežasčių. Anomalijų aptikimas gali padėti sužinoti kas lėmė statistinių rodiklių pokyčius. Tai yra labai aktualu, dėl didelės konkurencijos el. prekyboje. Bendrovės aptikusios anomalijas galėtų surinkti daugiau informacijos apie tą laikotarpį, pasikeitimus per jį ir sužinoti galimas rodiklio pasikeitimo priežastis. Tai leistų bendrovėms geriau pažinti savo pirkėjus ir kontroliuoti rodiklių pokyčius.

Šiame darbe bus tiriama „Little Data“ vieno iš klientų internetinės svetainės statistiniai duomenys 2014-07-01 – 2016-07-01 metų laikotarpiu. Tyrimo naujumą parodo tai, kad pasirinkti metodai dar nebuvo panaudoti anomalijoms elektroninėje prekyboje aptikti.

Tyrimo problema – rasti metodą, kuris yra tinkamiausias aptikti anomalijas remiantis elektroninės prekybos įvertinimais ir statistiniais duomenimis.

Tyrimo tikslas – pagrįsti anomalijų aptikimo elektroninėje prekyboje naudą ir rasti geriausią metodą joms aptikti.

Tyrimo uždaviniai:

1. Teoriškai pagrįsti anomalijų aptikimo naudą elektroninėje prekyboje;
2. Atlikti duomenų analizę;
3. Atlikti anomalijų nustatymo tyrimą naudojantis autoregresijos su slenkančiu vidurkiu (ARIMA);
4. Atlikti anomalijų aptikimo tyrimą naudojantis vektorinės autoregresijos (VAR) ;

5. Atlikti anomalijų aptikimo tyrimą naudojantis „AnomalyDetection“ R paketu;
6. Palyginti gautus rezultatus ir išrinkti geriausią metodą;

Šis tyrimas buvo atliktas naudojantis Microsoft „Word“ ir „RStudio“ programine įranga ir jos paketais.

1. LITERATŪROS APŽVALGA

Šioje darbo dalyje bus pateikta literatūros apžvalga apie anomalijų atpažinimo elektroninėje prekyboje naudą. Iš pradžių pateikta kaip elektroninė prekyba paveikė visą prekybą, kaip pakito ryšys tarp pardavėjo ir pirkėjo, kas paskatino ją atsirasti ir kokie veiksniai gali lemti pirkėjų požiūrį. Vienas iš būdų pagerinti elektroninės prekybos rezultatus yra anomalijų atpažinimas. Tai gali padėti išsiaiškinti staigius duomenų pokyčius, kas padėtų suprasti pirkėjų elgesio priežastis Todėl toliau dalyje taip pat bus aprašytos anomalijų rūšys, tyrimų metodų tipai bei iššūkiai, su kuriais susiduria tyrėjai norėdami jas atpažinti.

1.1. Elektroninė prekyba

M. Niranjanamurthy, N. Kavyashree, S. Jagannath, C. Dharmendra [1] savo straipsnyje teigia, kad elektroninė (dar vadinama internetinė) prekyba dažniausiai yra apibrėžiama kaip produkto pirkimas ar pardavimas per internetą. Patys autoriai mano, kad bet koks sandoris vykdomas tik per elektroninius prietaisus gali būti laikomas elektronine prekyba. A. Bakanauskas ir V. Liesionis [2] elektroninę prekybą apibrėžia kaip verslo operacijų atlikimą ir įmonės veiklos organizavimą, naudojantis informacinėmis technologijomis duomenų perdavimo tinklų aplinkoje. Skirtingai elektroninė prekyba yra apibrėžiama ir skitngų valstybių Ekonominio bendradarbiavimo ir plėtros organizacijos (EBPO) [3]. Įvairios verslo veiklos yra vykdomos internete – mažmeninė prekyba, bankininkystė, investavimas, nuoma. Elektroninę prekybą galima suskristyti į tris įprastas prekybos kategorijas: verslas verslui (angl. *Business to Business – B2B*), verslas klientui (angl. *Business to Consumer – B2C*) ir klientas klientui (angl. *Consumer to Consumer – C2C*). A. Bakanauskas ir V. Liesionis [2] išskiria šiuos pagrindinius el. prekybos tipus:

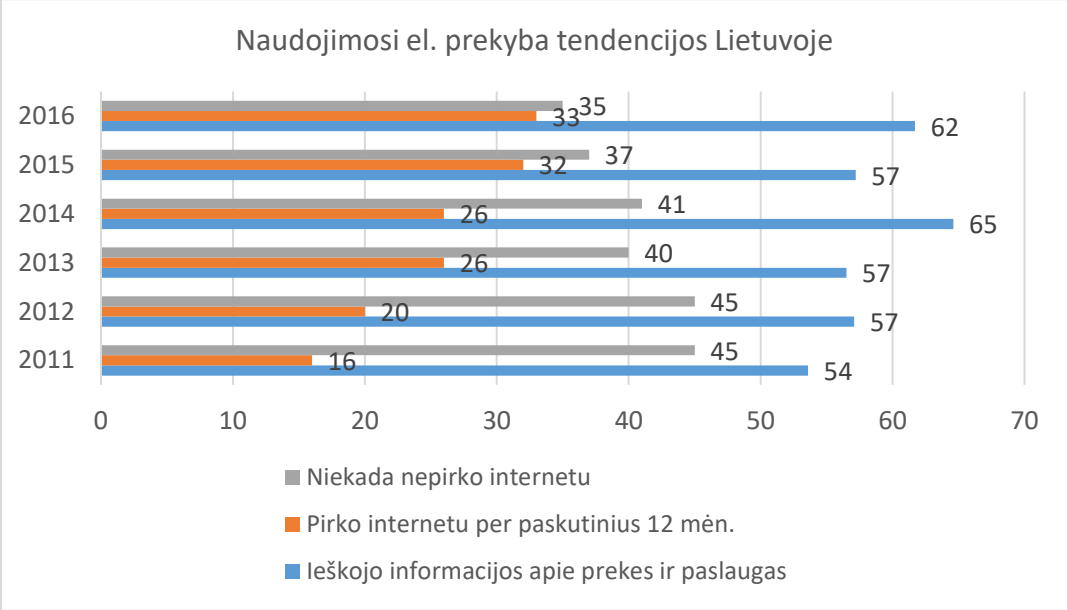
- El. prekyba prekėmis ir paslaugomis;
- Skaitmeninės informacijos pristatymas elektroninio ryšio priemonėmis;
- Elektroninė prekyba akcijomis;
- Elektroniniai aukcionai;
- Tiesioginis marketingas, skirtas vartotojams;
- Garantinės konsultacijomis elektroninėmis priemonėmis.

Tie patys autoriai išskyrė pagrindinius veiksnius, kurie lėmė el. prekybos atsiradimą ir sparčią jos plėtrą [2]. Jie plačiau pateikti 1.1. lent.:

1.1 lentelė. Veiksniai, kurie lėmė elektroninės prekybos atsiradimą (Šaltinis: [2])

Technologiniai veiksniai	• Asmeninių kompiuterių ir individualių technologijų paplitimas; • Interneto technologijos;
Ekonominiai veiksniai	• Verslo globalizacija; • Žemas startinio kapitalo slenkstis; • Standartinių prekių ir paslaugų paplitimas; • Finansavimo galimybės; • Elektroninio verslo prigimtis, gerai deranti su šiuolaikinėmis verslo savybėmis; • Tarpininkų mažėjimas ir jų vaidmens kitimas.
Socialiniai veiksniai	• Kvalifikacija, kuri reikalinga elektroninio verslo procesams kurti ir valdyti, yra gana lengvai įgyjama ir atnaujinama • Elektroninio verslo virtualumas ir tiesiogiškumas.

Elektroninė prekyba pakeitė pirkėjų patirtį apsipirkimo metu. Šiais laikais žmonėms nereikia važiuoti, eiti į fizinę parduotuvę norint įsigyti norimą daiktą ar paslaugą. Žmonės viską gali padaryti tiesiog sėdėdami namuose prie savo kompiuterio. Elektroninės prekybos pardavimų skaičiai yra labai sparčiai augantys, galimybė prisijungti prie el. parduotuvės platformos per daugybę skirtingų įrenginių iššaukė vartotojų prisirišimą prie technologijų[4]. Pagal naujausius Lietuvos Statistikos departamento [5] ir Europos statistikos departamento [6] duomenis (1.1. pav.) daugiau nei pusė žmonių, kurie naudojasi internetu ieškojo informacijos apie prekes ar paslaugas. Žmonių, kurie niekada nėra naudojęsi internetinės prekybos paslaugomis per paskutinius 5 metus sumažėjo 10%. Tuo tarpu žmonių, kurie pirko internetu per paskutinius 12 mėn. - išaugo daugiau negu dvigubai. Žinant kaip sparčiai keičiasi technologijos ir kaip jos tampa vis svarbesnė žmonių gyvenimo dalimi, galima neabejoti, kad greičiau ir geriau prisitaikę prie pokyčių verslai išliks ir padidins savo pardavimus, tuo tarpu kitos bendrovės, kurios yra lėtesnės ir ne tokios „lanksčios“ gali bankrutuoti ar prarasti savo pozicijas rinkoje.



1.1 pav. Naudojimosi el. prekyba tendencijos Lietuvoje 2011 – 2016 m. laikotarpiu. (šaltinis: [5], [6])

Pagal statistikos duomenis matome tendenciją, kad pirkėjų skaičius,pasinaudojusių internetinės prekybos paslaugomis, kiekvienais metais didėja. A. Bakanauskas ir V. Liesionis [2] savo knygoje pritaria, kad pardavimai ir paslaugos telefonu ir internetu vis didėja. Intenetas leidžia pardavėjams pasiūlyti savo paslaugas daugiau žmonių ir už mažesnes išlaidas.[7] Ypač ryškią tendenciją jie pastebi verslo verslui (B2B) sektoriuje.

Kaip pagrindinius kriterijus, kurie elektroninę prekybą padaro ypač patrauklią pirkėjo atžvilgiu A. Bakanauskas ir V. Liesionis [2] išskyrė tokius:

- **Vertė vartotojui.** Internete pirkėjus nuo konkurentų skiria tik keli mygtuko paspaudimai, jiems nereikia važiuoti į konkurentų parduotuvę, todėl prekės vertę lemia tokių veiksnių kombinacijos kaip prekės ypatumai, transakcijos kaina, aptarnavimas, rizikos laipsnis, išlaidos aptarnavimui.
- **Kaina.** Tai yra pats svarbiausias veiksnys pirkėjams, tačiau ne visiems. Aptarnavimo ar prekės gavimo laikas taip pat labai daug lemia, dėl šio veiksnio, dažnai pirkėjas yra pasiryžęs sumokėti daugiau. Kartais pirkėjas gali sumokėti daugiau ir jei jau yra anksčiau pirkęs iš šios internetinės parduotuvės ir pasitiki jos aptarnavimu.
- **Personalizavimas.** Tai leidžia padidinti vertę vartotojui. Didžiausios bendrovės siūlo pačiam pirkėjui pasirinkti prekės dizainą ir taip susikurti unikalų produktą.

- **Greitis.** Tai taip pat vienas iš svarbiausių veiksnių. Pirkėjai nori greitai rasti reikiamą informaciją ir pačią užsakytą prekę gauti jau sekančią ar dar kitą darbo dieną. Vartotojai nenori pildyti ilgų registracijos ar prekės išsiuntimo formų.
- **Patogumas.** Šis veiksnys ypač svarbus užimtiems žmonėms. Šie pirkėjai nori pirkti, sumokėti ir sulaukti savo produkto greitai ir jiems patogiu laiku.
- **Paprastumas.** Pirkėjai nenori ilgai ir painiai ieškoti informacijos apie prekes ar kaip jas užsisakyti. Tiek pačių prekių aprašymai, tiek pristatymo formos turi būti paprastos ir aiškios, jose turi būti įvedama tik būtina informacija.
- **Asmeniškumas.** Leidžia pirkėjui pasijusti svarbiu, net ir nebendraujant tiesiogiai su kompanijos atstovais. Kai kurie interneto puslapiai turi iššokančius pagalbos laukelius, kur internetinės svetainės atstovas realiu laiku stengiasi padėti pirkėjui išsirinkti norimą prekę ar palengvinti apsipirkimą, jei jam kyla kokių neaiškumų. Taip pat kai kurie internetiniai puslapiai, remiantis praeities pirkinių istoriją. siūlo asmeninius nuolaidų pasiūlymus vartotojams.
- **Galimybė neskubėti.** Žmogus internete nejaučia spaudimo greitai priimti sprendimo pirkti ar nepirkti prekės. Jis gali apžiūrinėti produktus ir įvertinti alternatyvas neribotą laiką, gali išeiti ir vėl grįžti į internetinę parduotuvę.
- **24/7.** Internetinės parduotuvės veikia visą parą. Dauguma žmonių negali apsipirkti darbo dienomis, darbo valandomis, todėl jie tai daro vakarais ar savaitgaliais. Tačiau retai žmonės nori savo poilsio laiką skirti apsilankymams, važinėjimams po parduotuves. Internete žmonės gali naršyti bet kuriuo metu.

Tačiau internetinė prekyba turi ir nemažai trūkumų. A. Bakanauskas ir V. Liesionis [2] kaip pagrindinius trūkumus įvardijo:

- **Fizinio kontakto nebuvimas su prekėmis**. Pirkėjas negali apžiūrėti ir išbandyti prekės prieš pirkdamas.
- **Prekės pristatymas užima laiko.** Būtinumas būti namuose pristatymo metu. Užsisakius prekę vartotojas jos iškart negauna. Paprastai užsisakius prekę reikia laukti kelias darbo dienas. Jei prekės nėra sandėlyje, pristatymas užtrunka dar ilgiau – iki kelių savaičių ar net mėnesio.
- **Prekės kokybės garantija.** Nors dauguma internetinių svetainių garantuoja kokybę, tačiau pasitaiko įvairių atveju, kai pirkėjas gavęs savo produktą nėra ją patenkintas, o tada jam reikia prekę išsiųsti atgal ir laukti grąžinamų pinigų arba naujos prekės.

- **Interneto ryšio sparta ir kokybė** gali riboti pirkėjo galimybes pasinaudoti visomis interneto parduotuvės turiniu.
- **Nėra asmeninio kontakto su pirkėju.** Pardavėjui yra sunkiau paveikti pirkėjo pasirinkimą ir sprendimą. Tačiau parduotuvės konsultantas gali pakonsultuoti pirkėją ir padėti jam išsirinkti prekę, kuri geriau atitiktų jo poreikius.
- **Išlaidos prekybos internetu sistemai.**
- **Išlaidos prekių siuntimui.**

H. Chen, R. H. L. Chiang and V. C. Storey [8] savo straipsnyje tvirtina, kad būtent internetinės prekybos bendruomenėse buvo daugiausiai vilčių dedama į didžiuosius duomenis (angl. *big data*) ir jų panaudojimą. Ryškūs rinkos pasikeitimai (išaugęs el. apsipirkimų skaičius) buvo pastebėti, kai tokie el. prekybos milžinai kaip Amazon ar eBay pakeitė savo el. prekybos platformas ir produktų rekomendavimo sistemas [8]. Priešingai negu tradicinių operacijų įrašai, kurie buvo surinkti iš įvairių sistemų 1980 metų laikais, duomenys, kuriuos el. prekybos sistemos surenka iš interneto yra mažiau struktūrizuoti ir dažnai turi naudingos informacijos apie klientų nuomonę ir įpročius. Socialinių tinklų analitikai naudoja vartotojų nuomonės, teksto ir sentimentų analizės technikas. Įvairios analitikos technikos taip pat naudojamos produktų rekomendacinėms sistemoms sukurti – asociacijos taisyklės, duomenų bazių segmentavimas ir klasterizavimas, anomalijų aptikimas ar diagramų tyryba. Visa tai padeda bendrovėms geriau suprasti savo pirkėjus. Suprasdami juos, el. prekybos bendrovės gali geriau atitikti jų keliamus lūkesčius, todėl toliau bus apžvelgta svarbiausi internetinės svetainės kokybės veiksniai.

1.2. Veiksniai lemiantys internetinės svetainės kokybę

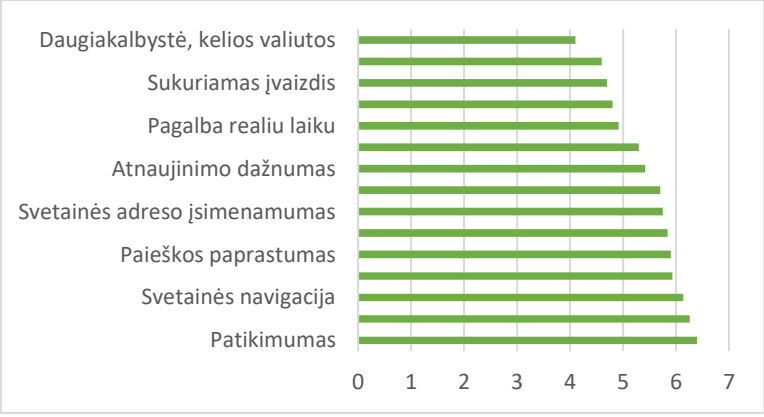
V. Davidavičienė ir J. Tolvaišas [9] savo sraipsnyje teigia, kad moksliniai tyrimai elektroninių paslaugų kokybės srityje yra fragmentiški. Tai lemia skirtingos verslo aplinkos sąlygos ir taikomi tyrimų metodai. Tačiau kiekvienas verslininkas norėtų žinoti kokie veiksniai lemia vartotojo požiūrį, įsitikinimus apie el. paslaugas ir internetinės svetainės kokybės suvokimą. Priklausomai nuo interneto svetainės tipo (komercinė, informacinė, mokomoji ir t. t.) yra pabrėžiami skirtingi interneto svetainės veiksniai jos kokybei įvertinti. V. Davidavičienė ir J. Tolvaišas [9] išskyrė pagrindinius el. puslapio kokybės įvertinimo veiksnius, kurie nepriklausomai nuo jo tipo:

- Paprastumas naudotis;

- Svetainės navigacija;
- Saugumo priemonės;
- Pagalba realiu laiku;
- Turinys;
- Dizainas;
- Paieškos paprastumas;
- Patikimumas;
- Pakrovimo laikas;
- Išsami kontaktinė informacija;
- Atnaujinimo dažnumas ir t.t.

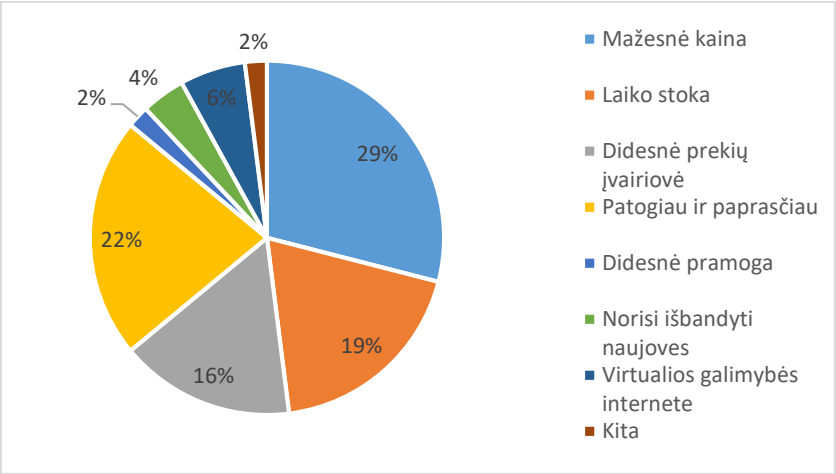
Pasak autorių dažniausiai, lygint kitus tyrimus, yra pabrėžiami šie veiksniai: svetainės navigacija, informacijos išdėstymas, paprastumas naudotis, pagalba realiu laiko momentu, patikimumas. Tačiau siekiant sudaryti elektroninės prekybos tipo svetainių kokybės vertinimą, būtina atsižvelgti į tam tikrus specifinius veiksnius kaip prekių paieškos lengvumas, užsakymo paprastumas, apmokėjimo procesas ir kt. V. Davidavičienė ir J. Tolvaišas [9] atliko apklausą Lietuvoje, kurios metu siekė nustatyti kokie kokybės veiksniai yra svarbiausi pirkėjams ir kokios priežastys juos skatina pirkti internetu.

Tarp veiksnių, kurie turi didžiausią svarbą el. prekybos interneto svetainės kokybei lietuviams galima įvardinti patikimumą, svetainės randamumą, navigaciją, turinį, paieškos paprastumą. Šių veiksnių bendri įvertinimai (skalėje 1-7) buvo didžiausi (1.2. pav.).



1.2 pav. Kokybės veiksnių svarba e. prekybos interneto svetainei (Šaltinis: [9])

Pagrindinės priežastys, kodėl pirkėjai renkasi apsipirkimą internete, apklausos duomenimis, yra mažesnė kaina (29 %), patogumas ir paprastumas (22 %) ir laiko stoka (19 %). Tyrimo rezultatuose nėra vienos priežasties, kuri dominuotų (ją būtų pasirinkę daugiau negu 50 % apklaustųjų). Tai aprodo, kad žmonės elektroninę prekybą renkasi dėl įvairių jos privalumų.



1.3 pav. Dažniausios naudojimosi e. prekyba priežastys (šaltinis: [9])

Žinant priežastis lemiančias pirkėjų apsisprendimą pirkti ar ne, kiekvienas elektronine prekyba užsiimantis asmuo gali tai pritaikyti savo verslui. Atlikti tyrimai gali padėti suprasti, stipriąsias ir silpnąsias internetinės svetainės puses. Tai atlikus galima atlikti pakeitimus. Norint įvertinti pasikeitimus galima naudoti el. prekybos vertinimo rodiklius. Šie rodikliai padeda įvertinti pirkėjo požiūrį.

1.3. El. Prekybos vertinimo rodikliai

Kiekviena įmonė renka informaciją apie savo pirkėjus, puslapio lankomumą ir kitus duomenis, kad vėliau galėtų iš jų išgauti vertingos informacijos. Didelės bendrovės pačios turi savo analitikos skyrių, o mažesnės dažniausiai pasinaudoja samdomų analitikų paslaugomis. Pati duomenų bazė yra nieko verta, jei bendrovė neturi žmogaus, kuris gali iš jos išgauti naudingos informacijos. Norint ją išgauti analitikams reikia atlikti daug darbo: ištraukti visą informaciją iš duomenų bazės, sutvarkyti duomenis (pašalinti tuščias reikšmes, netinkamus duomenis). Dažniausiai būtent duomenų tvarkymas

užima daugiausiai laiko. Nesutvarkius jų, atlikti tyrimai, sudaryti modeliai bus netikslūs ir neduos jokios (ar mažai) naudos.

Turėdami duomenis ir analitiką, kuris puikiai išmano savo darbą bendrovė gali geriau suprasti pirkėjų elgesį lemiančius veiksnius. Atlikus analizę galima palyginti skirtingų marketingo rūšių, kanalų rezultatus. Taip pat galima įvertinti lankytojų požiūrį į atliktus pasikeitimus. Tai gali padėti nustatyti kas skatina naujų pirkėjų atsiradimą ar esamų pirkėjų lojalumą.

G. R. Powell [10] išskyrė rodiklius į grupes pagal teikiamą informaciją. Pats autorius šiuos rodiklius apibūdina kaip sėkmės rodiklius (angl. *success metrics*), kurie parodo vartotojų reakciją į marketingo veiksmus. Sėkmės rodikliai yra suskirstomi į tokias kategorijas:

- **Finansinės sėkmės metrikos** – rodikliai, kuriuos galima gauti iš vidinių finansinių įmonės duomenų bazės. Tai pajamos, parduotų prekių ar produktų skaičius;
- **Tyrimais pagrįstos į pirkimus nukreiptos metrikos**. Šie rodikliai susistemina vartotojų požiūrį, ar parodo, ką vartotojai įsimena. Galime išskirti tokius rodiklius: prekės ženklo žinojimas, prekės ženklo vertė, pirkimo ketinimas, vartotojo ketinimas rekomenduoti produktą ar paslaugą;
- T**iesiogiai apskaičiuojamos preliminarios metrikos**. Vartotojų reagavimo rinkoje rodikliai. Pvz., paspaudimai, apsilankymai ar atsisiuntimai iš internetinės svetainės, klientų skaičius užregistruotas sistemoje;
- **Netiesioginės preliminarios metrikos**. Tai rodikliai nurodantys rinkos ypatumus, tačiau nebūtinai tiesiogiai apskaičiuojami;
- **Apskaičiuojamos metrikos**. Rodikliai apskaičiuojami pagal formules arba remiantis kitais rodikliais. Pvz., lojalumas, vartotojų išlaidų krepšelis, rinkos dalis, vartotojų gyvenimo vertė;

J. DeMers [11], L. Hasan, A. Morris ir S. Probets [12] nurodo rodiklius, kuriuos privalo naudoti kiekvienas, kuris nori tinkamai įvertinti savo internetinės veiklos efektyvumą:

1.2 lentelė. Svarbiausi elektroninės prekybos įvertinimo rodikliai (Šaltinis: [11],[12])

Rodiklis	Paaiškinimas/ svarba
Apsilankymų skaičius (angl. *Total Visits*)	Apsilankymų skaičiaus rodiklio vertinimas parodo bendrą vaizdą, kaip internetinis puslapis tvarkosi su vartotojų srautu. Jei pastebimas staigus kritimas, galima suprasti, kad reikia patikrinti savo rinkodaros kanalus ir išsiaiškinti to priežastis. Stabilioje bendrovėje apsilankymų skaičius turi augti vienodu tempu.
Sesijų skaičius, naujų sesijų skaičius (angl. *sessions, new sessions*)	Ši metrika parodo kiek interneto puslapio lankytojų yra naujų, o kiek lankosi joje jau kurį laiką. Tai puikus rodiklis patikrinti ar jūsų internetinis puslapis skatina lojalumą ir ar sugeba pritraukti naujų lankytojų. Tarkime, jei kardinaliai pakeitus puslapio struktūrą ar turinį, santykis tarp visų sesijų ir naujų sesijų nukrito, tai parodo, kad tinklalapis praranda efektyvumą išlaikyti nuolatinius lankytojus.
Kanalo srautas (angl. *Channel – Specific Traffic*)	Šis rodiklis yra labai svarbus, jei naudojama daug rinkodaros kanalų. Jis leidžia atrinkti kanalus, kurie yra geriausi ir pritraukia daugiausiai lankytojų.
Atmetimo rodiklis (angl. *Bounce rate*)	Atmetimo rodiklis parodo kiek lankytojų išeina iš tinklalapio tik jį atidarę, neapžiūrėję jo detaliau. Šis rodiklis gali parodyti ar tinklalapio struktūra ir turinys yra gerai išdėstytas (vartotojui draugiškas) ar lankytojams lengva rasti ko jie ieško.
Bendras konversijos rodiklis (angl. *Total Conversions*)	Tai pats svarbiausias rodiklis vertinant rinkodaros kanalų pelningumą. Mažas rodiklis gali reikšti prastą tinklalapio dizainą, blogus pasiūlymus ar tiesiog nesusidomėjusius lankytojus.
Klientų išlaikymo rodiklis (angl. *Customer Retention Rate*)	Rodiklis, kuris parodo kiek pirkėjų po vieno apsipirkimo sugrįžta į puslapį nusipirkti dar kartą. Mažas rodiklis gali parodyti prastą produkto ar aptarnavimo kokybę.
Kliento vertė (angl. *Customer Value)*	Šį rodiklį yra sunku apskaičiuoti, nes reikia numatyti viso kliento gyvavimo laikotarpio vertę. Tai gali padėti numatyti investicijos grąžą, numatyti bendrovės metinius tikslus.
Investicijų grąža (angl. *Return on Investment*)	Investicijų grąža yra pats svarbiausias rodiklis. Jis parodo kokį pelną atneša rinkodaros kompanija. Teigiamas rodiklis parodo, kad rinkodaros strategija yra efektyvi, neigiama – reikia rimtų pokyčių joje.

Dauguma šių rodiklių yra plačiai žinomi ir naudojami, didžiausias jų privalumas tai, kad jie yra skaitiniai – tai leidžia juos palyginti. Turėdami ilgo laikotarpio rodiklius ir analitiką, kuris galėtų juos išanalizuoti, galima geriau suprasti kokios priežastys lemia pirkėjų veiksmus. Tačiau ne visada yra paprasta suprasti rodiklių svyravimų priežastis. Staigūs pokyčiai rodikliuose gali būti vertingi bendrovės vadovams. Tokie svyravimai, kurių negalima paaiškinti vadinami anomalijomis arba

išskirtimis. Aptikus anomalijas ir gavus daugiau duomenų būtų galima išsiaiškinti šuolių priežastis, o tai padėtų juos suvaldyti.

1.4. Anomalijos internetinėje prekyboje

Šioje darbo dalyje aprašyta kaip įvairūs šaltiniai apibrėžia anomaliją. Taip pat pateikta į kokius metodus yra skirtomi anomalijų atpažinimo būdai ir su kokiais iššūkiais susiduriama aptinkant anomalijas.

1.4.1. Anomalijos samprata

Statistikos bendruomenė jau 19 amžiuje pradėjo tyrinėti kaip surasti anomalijas duomenyse (Edgeworth 1887). Per tą laiką įvairios anomalijų aptikimo technikos buvo sukurtos įvairių sričių bendruomenių. Dauguma iš jų buvo sukurtos būtent jų sričių panaudojimui, tačiau yra ir tokių, kurios yra bendrinės ir tinka įvairių sričių anomalijoms aptikti.

Terminų žodynas anomaliją apibrėžia kaip netaisyklingumą, nukrypimą nuo normos, nuo dėsningumo. Tačiau pats terminas anomalija yra vartojamas ne vienoje srityje, todėl norint jį apibrėžti taisyklingai matematinėje srityje, galima būtų remtis užsienio autorių straipsniais, tyrimais. V. Chandola, A. Banerjee ir V. Kumar [13] savo pranešime apie anomalijų atpažinimą patį reiškinį apibūdina būtent kaip matematinės srities terminą. Anomalijomis jie vadina tokius šablonus, kurie neatitinka numatytų reikalavimų, elgesio. Anomalijos taip pat dažnai yra vadinamos išskirtimi, nesiderinančiu stebiniu, išimtimi, aberacija, netikėtumu, savitumu.

V. J. Hodge ir Austin, J. [14] anomalija ar išskirtimi vadina stebinį, kuris gerokai nukrypęs nuo kitų narių imtyje. Savo tyrime jis taip pat pateikia ir šiek tiek kitokį išskirties apibrėžimą: Tai stebinys (ar poaibis stebinių), kurie nesuderinami su likusia duomenų aibe.

Išskirčių radimas yra labai svarbi dalis, norint sudaryti tikslų ir gerą modelį savo duomenims. M. R. Smith ir T. Martinez [15] teigia, kad „triukšmas“ ir išskirtys duomenyse atsiranda dėl tipografinių (duomenų suvedimo) ar matavimo klaidų. „Triukšmas“ yra svarbus todėl, kad šie duomenys gali paveikti pagal sistemos mokymosi algoritmą sukurtą modelį. Algoritmo gebėjimas susidoroti su šiomis anomalijomis lemia modelio gerumą, tikslumą.

Anomalijų atpažinimo metodai yra naudojami labai plačiai. R. J. Hyndman, E. Wang ir N. Laptev [16] teigia, kad vis daugiau bendrovių renka didelės apimties duomenis ir siekia juose įžvelgti neįprastas ar anomalijų laiko eilutes. V. J. Hodge ir J. Austin [15] savo tyrime apie anomalijų suradimo technikas išdėstė daugybę sričių, kur šie tyrimai yra naudingi:

- Apgavysčių aptikimas – aptikti apgaulingas programas kredito kortelėms, valstybės lėšoms ar atpažinti apgaulingai naudojamas kredito korteles, mobiliuosius telefonus.
- Paskolos gavimo apdorojimas – atpažinti apgaulingas paraiškas ir potencialius problemiškus klientus.
- Įsibrovimų aptikimui – aptikti neteisėtus prisijungimus prie kompiuterio tinklo.
- Veiklos stebėjimas – aptikti mobiliųjų telefonų apgavystės atvejus, stebint telefono veikslą ar įtartinus sandorius nešališkose rinkose.
- Tinklo veikla – prižiūrėti kompiuterio tinklo veiklą, naudinga rasti tinklo „butelio kakliuką“.
- Gedimo diagnostika – prižiūrėti procesus, surasti gedimus varikliuose, generatoriuose.
- Struktūrinių gedimų aptikimas – prižiūrėti gamybos linijas, aptikti produkcijos gamybos defektų priežastis.
- Palydovo nuotraukų analizė – nustatyti naujas ar netinkamas ypatybes ir jas pakeisti.
- Judėjimo segmentavimas – aptikti paveiksliuko ypatybes nepriklausomai nuo jo fono.
- Laiko eilutės stebėjimas – stebėti kritinio saugumo programas, pvz., gręžimą, greitąjį frezavimą.
- Sveikatos būklės stebėjimas – širdies dažnio stebėjimas.
- Farmaciniai tyrimai – naujų molekulinių struktūrų aptikimui.
- Aptikti nenumatytus įrašus duomenų bazėje – duomenų tyrybai, aptikti klaidas, sukčiavimą ar galiojančius bet nenumatytus įrašus.

Nors autorius nepabrėžė mums aktualiausios srities – el. prekybos srauto stebėjimo, tačiau galime įsitikinti, kad išskirčių aptikimas yra labai plačiai naudojamas ir yra daugybė būdų jiems aptikti.

1.4.2. Anomalijų atpažinimo metodai

V. J. Hodge ir J. Austin [14], V. Chandola ir kt. [13] anomalijų aptikimo metodus suskristo į 3 pagrindines grupes:

1. *Išmokyti anomalijų aptikimo metodai*. Tai metodai, kai modeliai yra mokomi naudojantis mokymui skirtu duomenų rinkiniu, pagal kurį modelis išmoksta, kokie duomenys yra normalūs, o kokie yra anomalijos. Paprastai tokiu atveju yra sukuriamas prognozavimo modelis, kuris suskirsto duomenis į 2 grupes: normalūs ir anomalijos. Visi nauji duomenys yra palyginami su tais, kuriais modelis buvo mokytas ir priskiria juos į vieną iš grupių. Tiesa, yra 2 didelės problemos, kurios iškyla naudojant išmokytus anomalijos aptikimo metodus. Pirma, anomalijų atvejų paprastai yra žymiai mažiau nei normalių atvejų. Duomenų disbalansas gali būti svarbi problema sukuriant gerą modelį. Antra problema yra, tinkamas ir tikslus anomalijų grupės bruožų identifikavimas. Yra sukurti keli metodai, kurie siūlo įterpti dirbtines anomalijas mokymo rinkinio duomenyse, kad modelis lengviau aptiktų išskirtis.

2. *Pusiau išmokyti anomalijos aptikimo metodai.* Tai metodai, kai mokymo rinkinys yra naudojamas vienam iš grupių bruožams nustatyti (pvz., normaliems duomenims). Šie metodai mokymo metu naudojant duomenų rinkinį išmoksta atpažinti normalius duomenis, o pateikus testavimo duomenis, modelis neįprastus duomenis priskiria prie anomalijų.

3. *Neišmokinti anomalijų aptikimo metodai.* Tai metodai, kai modelio nereikia mokinti atskiru duomenų rinkiniu. Šie metodai padaro be sąlygišką prielaidą, kad rinkinyje yra gerokai daugiau normalių duomenų negu anomalijų. Jei ši prielaida yra neteisinga – šie metodai prastai atpažįsta išskirtis.

Dauguma pusiau išmokytų metodų gali būti naudojami ir kaip išmokyti. Tokiu atveju nereikia sudaryti mokymo rinkinio. Toks pritaikymas padaro prielaidą, kad testavimo rinkinyje yra labai mažai anomalijų ir todėl modelio išmokymas būtų labai sudėtingas.

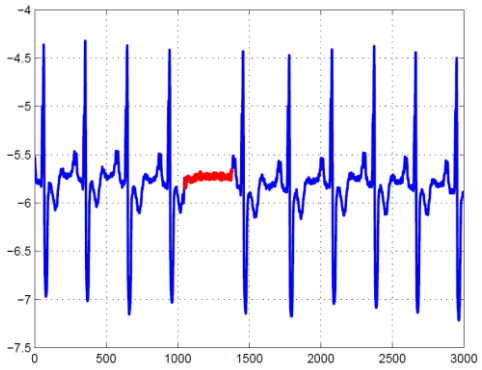
Tiesa, pačios išskirtys taip pat būna nevienodos, kartais tai yra tik vienas stebinys, kartais laiko eilutėje keli stebiniai, esantys šalia vienas kito išsiskiria iš visų likusių. Todėl norint pasirinkti tinkamą metodą, reikia atkreipti dėmesį kokių anomalijų ieškoma ar galbūt, žinoma kokios jos yra duomenyse. V.Chandola ir kt.[13] savo tyrime išskyrė tokias rūšis:

1. **Anomalijos taškai** *(angl. Point Anomalies).* Tai anomalijos rūšis, kai vienas taškas išsiskiria iš likusių duomenų. Tai pati paprasčiausia išskirties rūšis. Dažniausiai tyrimai yra atliekami būtent tokioms anomalijoms atpažinti.

2. **Kontekstinė anomalija (**angl. *Contextual Anomalies)*. Tai anomalijos rūšis, kai duomenys yra laikomi išskirtimi esant tam tikai situacijai. Pati situacija turi būti aiškiai parodyta duomenų rinkinio struktūroje ir turi būti įvardinta iškeliant problemą. Paprasčiausias pavyzdys galėtų būti oro

temperatūra. Žiemą yra įprasta neigiama ar artima nuliui oro temperatūra. Tuo tarpu vasarą - tokia temperatūra būtų išskirtis, nes paprastai vasarą oro temperatūra yra tarp 15 - 25 laipsnių.

3. **Kolektyvinės anomalijos** (angl. *Collective Anomalies*). Jei grupė susijusių duomenų yra anomalijos palyginus jas su likusiais duomenimis, tai tokia grupė yra vadinama kolektyvinė anomalija. Atskiri duomenų taškai kolektyvinėje anomalijoje gali būti ir ne išskirtys, bet jie visi kartu sudaro anomaliją. Kaip pavyzdį galime matyti 1.4. pav. pateiktus elektrokardiogramos duomenis, kur raudonai yra pažymėta kolektyvinė anomalija. Kaip matome, toje dalyje ta pati maža reikšmė laikosi neįprastą laiko tarpą. Tai puikus pavyzdys, kai patys taškai atskirai nėra anomalijos.



1.4 pav. Kolektyvinės anomalija. Elektrokardiograma (šaltinis: [13])

Atliekant tyrimą ir pasirinkus metodą yra svarbu žinoti kaip norima, kad rezultatai būtų pateikti. Dažniausiai metodai leidžia pasirinkti vieną iš šių rezultatų rūšių:

- Balas (angl. *Score*). Balų skyrimo metodu, modelis testavimo duomenims skiria balą remdamasis tuo kaip tikėtina, kad taškas yra anomalija. Tokio metodo rezultatas yra pagal reikšmę išdėstytų anomalijų sąrašas. Modelis gali arba pats automatiškai parinkti balo reikšmę, nuo kurio dėmuo laikomas anomalija. Taip pat galima pačiam pasirinkti ribą, nuo kurios taškas yra priskiriamas anomalijai.
- Etiketė (angl. *Label*). Šiuo būdu modelis priskiria taškams etiketes, t. y. taškas yra priskiriamas prie anomalijų arba normalių duomenų.

Rezultatai paremti balų skyrimu leidžia žmogui pasirinkti slenkstį, kurį viršijus taškas priskiriamas anomalijai, tai leidžia atrinkti duomenis, kurie yra panašiausi į anomalijas. Tuo tarpu

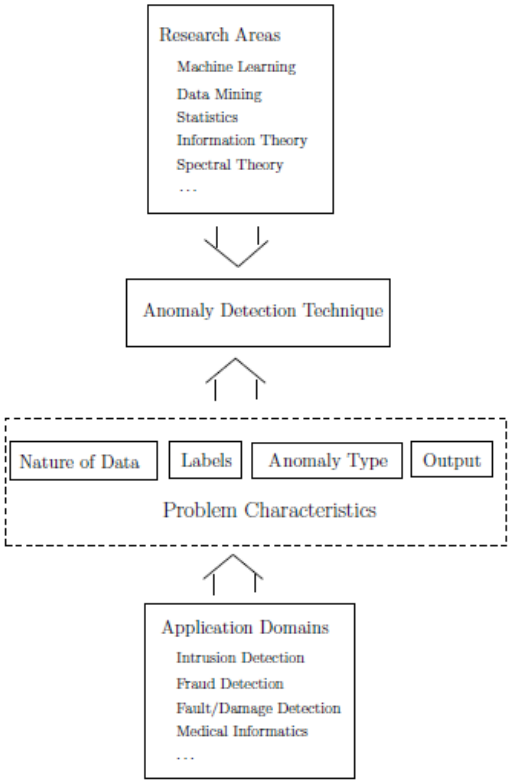
etiketės rezultatas neleidžia to padaryti iš karto, tačiau dauguma metodų turi papildomus parametrus, kurie leidžia pasirinkti ribą.

1.4.3. Anomalijų tyrimo iššūkiai

Teorijoje anomalija apibūdinama, kaip neįprasta duomenų dalis ar tik vienas taškas, kuris išsiskiria iš numatomų normaliųjų. Anomalijos aptikimas turi nurodyti koks yra normalus stebinys ir koks stebinys tokiam nepriklauso ir yra priskiriamas išskirčiai. Bet yra faktorių, kurie šią paprastą užduotį padaro ne tokią paprastą:

- Apibūdinti normalių duomenų ribas, kurios tiksliai apibrėžtų jų savybes yra labai sudėtinga. Dažniausiai ribos tarp normalių ir anomalijai būdingų savybių nustatyti tiksliai yra neįmanoma. Dažnai anomalijai priskirtas stebinys, kuris yra arti ribos iš tikrųjų yra normalus ir atvirkščiai.
- Kai anomalija yra piktybinių veiksmų pasekmė, tai priešinga pusė dažnai prisitaiko prie to ir ją pakeičia į normalų stebinį, tai padaro normalių stebinių savybių apibrėžimą dar sudėtingesnį.
- Daugumoje sričių normalių stebinių savybės vis keičiasi ir dabartinės savybės gali nebetikti šiam tipui ateityje (gali būti priskiriama anomalijai).
- Anomalijos supratimas gali skirtis priklausomai nuo srities. Pavyzdžiui, medicinoje mažas nukrypimas nuo normalios gali būti laikomas anomalija (-pvz., kūno temperatūros svyravimai), kai tuo tarpu toks pats pokytis akcijų rinkos srityje (-akcijų vertės pasikeitimas) gali būti laikomas normaliu. Dėl šios priežasties to pačio metodo naudojimas keliose srityse gali būti prastas pasirinkimas.
- Didelis duomenų disbalansas (didelė dalis duomenų yra normalūs). Labai maža dalis anomalijų duomenyse apsunkina jų savybių apibrėžimą modeliuose.
- Dažnai duomenyse esantis „triukšmas“ yra labai panašus į anomalijas, todėl yra sudėtinga jį surasti ir pašalinti.

Dėl visų anksčiau išvardintų priežasčių tinkamai ir tiksliai surasti anomalijas yra sudėtinga užduotis. Dauguma esančių anomalijų aptikimo būdų yra skirti tam tikrai suformuluotai problemai. Pati problema yra suformuluota daugybės faktorių, tokių kaip pačių duomenų, kiek iš jų yra priskirti normaliems ar anomalijoms, kokio tipo anomalijas norima aptikti ir t. t. Dažniausiai šie veiksniai yra numatomi pagal sritį, kuriose siekiama atpažinti anomalijas.



1.5 pav. Pagrindinės sudedamosios dalys, susijusios su anomalijos atpažinimo technikomis (šaltinis: [13])

Tyrėjai pritaikė sąvokas iš skirtingų sričių (statistikos, sistemos mokymosi, duomenų tyrybos, informatikos) ir pritaikė apibrėžti anomalijų aptikimo problemai. 1.5. pav. parodytos pagrindinės sudedamosios dalys, susijusios su anomalijos atpažinimo technikomis. Galima matyti, kad yra 3 pagrindinės dalys anomalijų atpažinimo technikai: tyrimo sritis (angl. *Research area*), taikymo sritis (angl. *Application domains*) ir problemos savybės (angl. *Problem characteristics*).

1.5. Skyriaus apibrendrinimas

Atlikus literatūros apžvalgą, galima įsitikinti, kad elektroninė prekyba tampa labai svarbia prekybos dalimi. Lietuva taip pat neatsilieka ir gyventojų, kurie apsipirkinėja interneto pagalba vis daugėja. Tačiau elektroninės parduotuvės vis dar tiksliai negali prognozuoti vartotojų veiksmų. Vis dar nėra žinoma kaip galima vartotojus „priversti“ pirkti. Anomalijų aptikimas įvertinimuose ir statistiniuose rodikliuose padėtų išaiškinti žymius statistikos pasikeitimus. Tai gali padėti geriau suprasti pirkėjus. Toliau darbe bus pateikti pasirinktų metodų tirti anomalijas teorinė dalis.

2. MEDŽIAGOS IR TYRIMŲ METODAI

Šioje darbo dalyje pateikta tyrimo dalyje naudotos medžiagos ir metodų teorinė dalis. Pradžioje pateikta duomenų tipo – laiko eilutės teorinė dalis. Po jos – duomenų charakteristiką padėsiantis išaiškinti metodai. Ištyrus duomenis, reikia pritaikyti pasirinktus modelius. Siekiant surasti geriausią metodą, kuris atpažintų anomalijas buvo išbandyta daug modelių: Neuroniniai tinklai (angl. *neural network*), k-artimiausio kaimyno(angl. *k-nearest neighbor*), Bajeso, atraminių vektorių (angl. *Support vector machine*), Sprendimų medžio (angl. *Decision Tree*), vietinių anomalijų faktoriaus (angl. *Local Outlier Factor*) metodai. Deja, visi metodai labai prastai atpažino anomalijas. Todėl ieškoti kitokių metodų. Šiame skyriuje bus pateikta teorinė dalis autoregresijos su slenkančių vidurkiu (ARIMA), vektorinės autoregresijos (VAR) ir „AnomalyDetection“ R paketo modelių dalis. Rezultatams įvertinti buvo naudojama sumaišymo matrica (angl. *confusion matrix*) ir jos įvertinimai.

2.1. Laiko eilutė

Laiko eilute galima vadinti tam tikro atsitiktinio kintamojo (dydžio) reikšmes stebimas keičiantis laikui (reikšmės stebimos kas vienodą laiko tarpą). Pavyzdžiui, šalies arba regiono gyventojų skaičiaus pokytis, tam tikrų visuomenės sluoksnių socialinio aktyvumo dinamika, žemės ūkio plėtros rodiklių kitimas ir pan. Laiko eilučių analize vadiname tokią analizę, kurioje tiriami duomenų taškai kintantys laike. Analizės tikslas yra nustatyti, kokį efektą kintamajam Y turės kintamojo X pasikeitimas per tam tikrą laiką [17].

Tiriant laiko eilutes, priimta, kad žinomos reikšmės tam tikrais laiko momentais , o visos reikšmės fiksuojamos vienodais laiko intervalais ($t_{i+1} - t_i = \Delta t$)

Jeigu yra stebimas vieno rodiklio reikšmių pokytis, tokią laiko eilutę vadiname vienmate, o stebint daugiau negu vieno rodiklių reikšmių kitimą, tokia eilutė vadinama *daugialype*.

Lyginant laiko eilutės ir atsitiktinio duomenų rinkinio stebinius, naudotus regresinėje analizėje, galima išskirti tokius skirtumus:

1) laiko eilutės stebiniai yra priklausomi tarpusavyje;

2) laiko eilutės stebiniai yra nevienodai pasiskirstę laiko ašyje:

$$P\{Z(t_1 < Z\} \neq P\{Z(t_2) < Z\}, \text{ kai } t_1 \neq t_2, \quad (1)$$

čia *Z(t)* - tolydi laiko eilutė, kai laikas yra nenutrūkstamas. Jei laikas diskretus – eilutę galima vadinti diskrečia. Ji žymima Z_t. Nagrinėjant įvairius procesus, dažniausiai yra analizuojamos diskrečios laiko eilutės. Apibrėžiant diskrečias eilutes yra svarbu užfiksuoti ne tik laiko intervalą Δt bei stebėjimų skaičių *n*, bet ir pradinį laiko momentą t_0. Norint analizę padaryti paprastesne dažnai $t_0 = 0$.

Taip pat laiko eilutės gali būti *momentinės* (pagamintos produkcijos skaičius) ir *intervalinės* (per dieną parduotos produkcijos apimtys). Laiko eilučių analizė atliekama dviem aspektais: dažnumo srities ir laiko srities.

Jei modelis yra atliekamas naudojant dažnumo srities metodus, tai laiko eilutė aprašoma nepriklausomų, kosinuso ir sinuso su skirtingomis amplitudėmis, dėmenų suma. Šio modelio matematinė formulė yra tokia:

$$Z_t = \mu + \sum_j \left[Y_j \cos(2\pi\omega_j t) + X_j \sin(2\pi\omega_j t \right], \quad (2)$$

čia Y_j, X_j - nekoreliuoti atsitiktiniai kintamieji su nuliniu vidurkiu ir dispersija $\sigma^2(\omega_j)$.

Dažniai $\omega_1, \omega_2, \ldots$ yra nustatomi mažais intervalais $\Delta\omega$. Tokia analizė dar yra vadinama spektrine analize. Dažniausiai ji yra naudojama elektrotechnikos, radioelektronikos, telekomunikacijų moksluose.

Analizuojant laiko eilutes reikia išspręsti 3 pagrindinius uždavinius:

1. *Identifikacijos*, t. y. modelio parametrų statistinis įvertinimas;
2. *Verifikacijos*, t. y. sudaryto modelio adekvatumo patikrinimas;
3. *Prognozavimo*, t. y. laiko eilutės reikšmių laiko momentais nustatymas, kai $l > n$.

Modelio *identifikacija*:

- Laiko eilutės duomenų atvaizdavimas ir tinkamos transformacijos parinkimas (duomenis paversti stacionariais);
- ACF ir PACF analizė;
- Modelio parametrų įvertinimas – įvertinami parametrai, naudojant maksimalaus tikėtinumo metodą.

Modelio *verifikacija* (patvirtinimas*)*:

- Išmokius modelį yra patikrinamas modelio tikslumas su kitais duomenimis. Šiuo būdu yra tikrinama ar modelis nėra „permokytas“. Kitaip sakant, tikrinama ar sudarytas modelis tinkamas naudoti su bet kokiais duomenimis.

Modelio *prognozavimas*:

- Naudojamas atskiras duomenų rinkinys įvertinti modelio prognozavimo tikslumą. Prognozavimas įvertinamas palyginus prognozavimo ir realias duomenų reikšmes;

2.1.1. Laiko eilutės stacionarumas

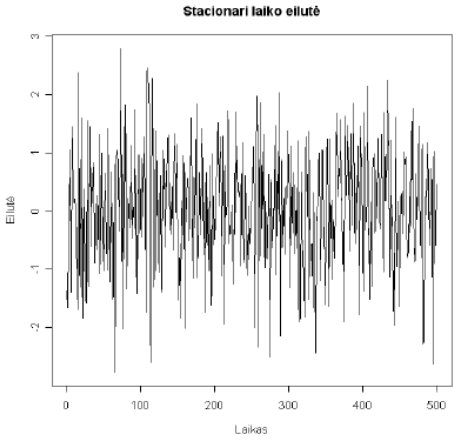
Norint duomenis pritaikyti dažniausiai naudojamiems modeliams, viena iš būtinų sąlygų yra stacionarumas. Dažniausiai realių įvykių ir realūs (nesugeneruoti) duomenys yra nestacionarūs. Todėl norint sudaryti tikslų modelį yra būtina ją transformuoti į stacionarią. Sakoma, kad laiko eilutė yra stacionari siaurąja prasme, jei jos daugiamačiai pasiskirstymai nepriklauso nuo poslinkio laike:

$$F_{t_1 \ldots\ldots t_k}(\cdot) = F_{t_1 \ldots\ldots t_k+\tau}(\cdot), \forall t_1, \ldots\ldots, t_k, \tau \in T. \quad (3)$$

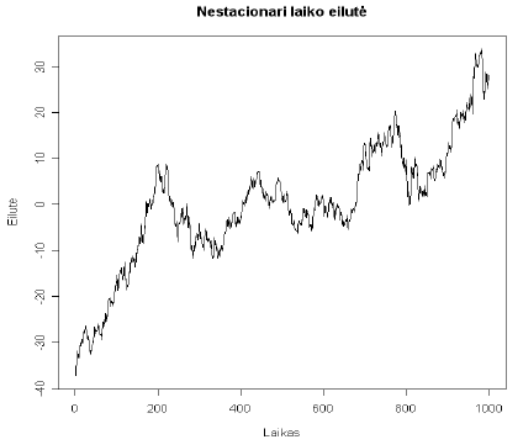
Laiko eilučių analizėje dažniausiai yra naudojamas stacionarumo plačiąja prasme apibrėžimas. Procesas Y_t yra stacionarus plačiąja prasme, jei:

1. $EY_t^2 < \infty, \forall\, t \in T$;
2. $EY_t < EY_0\,, \forall\, t \in T$;
3. $\mathrm{cov}(t,s) = \mathrm{cov}\,(t+h, s+h), \forall\, t,s,h \in T$.

Dažniausiai nustatyti ar duomenys yra stacionarūs ar ne galima ir iš grafinio eilutės vaizdo:



2.1 pav. Stacionari laiko eilutė



2.2 pav. Nestacionari laiko eilutė

Jei duomenyse yra trendas per visą laikotarpį, t. y. laiko eilutė per visą laiką nuolat kyla ar krinta – tokia laiko eilutė yra nestacionari (2.2. pav.).

Laiko eilutės stacionarumą galima nustatyti remianatis specialiais tam skirtais testais. Yra nemažai testų, kurie tinka šią užduotį atlikti. Šio tyrimo atveju buvo naudoti Box – Ljung ir Augmented Dickey-Fuller testai [18].

Box – Ljung testas

Box – Ljung testas yra naudojamas patikrinti liekanų nepriklausomumui. Jis patikrina ar kuri nors iš laiko eilutės autokoreliacijų grupės skiriasi nuo nulio. Box – Ljung testo hipotezė:

$$H_0: \rho_1 = \cdots = \rho_m = 0 \ (liekanos\ nekoreliuotos) \tag{15}$$

$$H_a: \exists_i \leq m: \rho_i \neq 0 \ (liekanos\ koreliuotos) \tag{16}$$

Testo statistika:

$$Q = n(n+2) \sum_{j=1}^{h} \frac{\widehat{\rho_j^2}}{n-j} \tag{17}$$

čia *n* -imties dydis, ρ_j – vėlavimų *j* autokoreliacija, *h* – skaičius tikrinamų lagų.

Ši statistika yra lyginama su χ – kvadrato su *h* laisvės laipsnių ir *a* reikšmingumo lygmeniu skirstiniu.

Augmented Dickey-Fuller testas

Dar vienas testas skirtas patikrinti duomenų stacionarumą. Jis yra vienas iš populiariausių testų. Vienetinė šaknis yra laiko eilutės autoregresijos parametras, kuris yra lygus 1. Jei laiko eilutės vienetinė šaknis yra lygi 1, tai ji yra laikoma nestacionaria. Tarkime, kad proceso liekanos yra AR procesas, o AR(*p*) yra X_t. ADF testo hipotezė:

$$H_0: procesas\ X_t\ turi\ vienetinę\ šaknį \tag{18}$$

$$H_a: procesas\ X_t\ yra\ stacionarus \tag{19}$$

AR procesą išreiškiame tokiu pavidalu:

$$\Delta X_t = (\propto -1)X_{t-1} + \sum_{i=2}^{p} \beta_i\, \Delta X_{t-i+1} + \ a_t \tag{20}$$

čia *p* – lagų skaičius, o a_t – baltasis triukšmas. $\propto = \ \sum_{j=1}^{p} \varphi_j\,, o\ \beta_i = \ \sum_{j=i}^{p} \varphi_j\,, i = 2, \ldots., p$

Pažymėjus $1 - \alpha = \Upsilon$. Vienetinės šaknies testas yra vykdomas nulinei hipotezei $\Upsilon = 0$ su stacionarumo alternatyva, kad $\Upsilon < 0$. Testo statistika, kuri yra

$$DF_t = \frac{\widehat{\gamma}}{SE(\widehat{\gamma})} \tag{21}$$

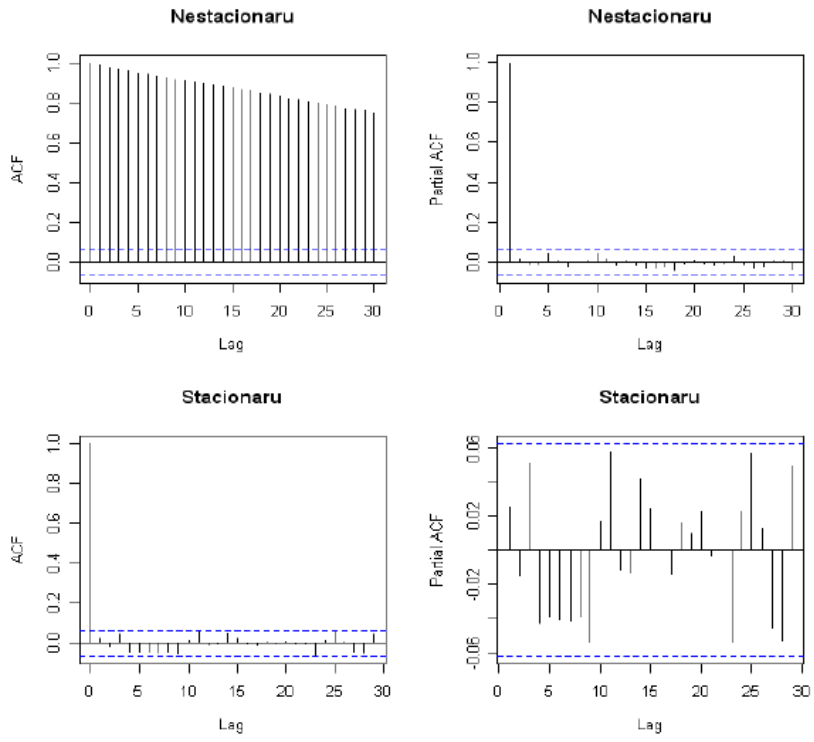
Lyginama su atitinkama Dickey – Fuller testo, taikomo autoregresiniam modeliui, kritine reikšme. Čia *SE* (Υ) – dydžio standartinė paklaida.

Jei nulinė hipotezė yra priimama, daroma išvada, kad X_t nėra stacionarus, ir pereinama prie tokio paties $(1 - L)\ X_t$ nagrinėjimo. Jei gaunama išvada, kad ir $(1 - L)\ X_t$ nestacionarus, tiriama antros eilės pokyčiai ir t. t.

2.1.2. Laiko eilutės autokoreliacija ir dalinė autokoreliacija

Kitas populiarus būdas laiko eilutės stacionarumui patikrinti yra autokoreliacijos grafikas. Autokoreliacija (ACF) yra matematinė panašumo laipsnio tarp laiko eilutės ir tos pačios eilutės, pastumtos laike, išraiška. Dalinė eilutės $\{Y_t\}$ autokoreliacija (PACF) yra koreliacija tarp Y_0 ir Y_k pašalinus tiesinę $Y_1, \ldots, Y_{k-1}$ regresiją.

Jei laiko eilutė yra stacionari – ACF grafiko reikšmės yra artimos nuliui ar neviršija numatytos ribos. Jei yra daug reikšmių, kurios viršija slenkstinę ribą, tokia laiko eilutė yra laikoma nestacionaria. Tuo tarpu dalinės autokoreliacijos grafikai leidžia numatyti duomenų sezoniškumą. Nestacionarios laiko eilutės dalinės autokoreliacijos grafiko reikšmės neviršija slenkstinės ribos. Tačiau stacionarios – viršija (2.3. pav.). ACF ir PACF grafikai padeda nuspręsti kokią įtaką dabartinei rodiklio reikšmei turi praeityje buvusios reikšmės. Pagal jų kitimą yra nustatomi autoregresijos ir slenkančio vidurkio eilės parametrai p ir q.



2.3 pav. Stacionarios ir nestacionarios laiko eilučių autokoreliacijos ir dalinės autokoreliacijos grafikai

2.2. Vektorinės autoregresijos modelis

Vektorinės autoregresijos modeliai (VAR) pradėti kurti nuo 1980 m. C.A.Sims [19]. Iki to laikotarpio laiko eilučių analizė apsiribodavo tuo, kad pradiniai duomenys visada turėdavo būti transformuojami siekiant padaryti juos stacionariais, o tai atlikus, buvo sudaromi ARMA modeliai. Dėl šio proceso buvo galima prarastį dalį naudingos informacijos, duomenų. Kadangi tiriant keletą laiko eilučių reikia įvertinti jų tarpusavio priklausomybę, buvo pasitelkiamas dinaminių ekonometrinių lygčių sistemos modelis. Naudojant jį reikėjo atlikti dvi procedūras:

- Visus kintamuosius reikia griežtai suskirstyti į dvi kategorijas: endogeninius ir egzogeninius;
- Siekiant modelį nustatyti - reikia pasirinkti parametrus ir juos įvertinti.

Sims nustatė, kad naudojant šias dvi procedūras tenka daryti konfliktinius sprendimus, ir pasiūlė alternatyvų VAR modelį [19],[20]. Šis modelis yra daugiamatės laiko eilutės autoregresijos modelio apibendrinimas. VAR modelio privalumas yra tai, kad jį įvertinti galima klasikiniu mažiausių kvadratų metodu (MKM). Taip pat, šiame modelyje *a priori* nereikia visų kintamųjų suskirstyti į dvi grupes (endogeninius ir egzogeninius), o tai yra svarbu, kai nėra iki galo aiški priežastinė priklausomybė. Vektorinės regresijos modelis yra pavadintas taip todėl, kad autoregresijos terminas yra vartojamas dėl laginių kintamųjų reikšmių, o vektorinis – nes aprašomas laiko eilučių vektorius.

Žinomiausias Vektorinės autoregresijos modelis buvo sukurtas investicinės kompanijos „JP Morgan“. Šis modelis buvo pavadintas „RiskMetrics“. Po to, kai 1994 metų lapkritį „JP Morgan“ šią sistemą padarė visiems prieinamą, rizikos valdymas, naudojant VAR, tapo labai populiarus tarp finansų rinkų dalyvių. Dabar bankai, draudimo kompanijos, pensijų fondai ir nefinansinio pobūdžio kompanijos naudoja VAR finansinei ir kitokiai rizikai valdyti. Pastaruoju metu šiais modeliais yra valdomos ne tik rinkos rizikos, tačiau ir likvidumo, kredito ir pinigų srautų rizikos.

Paprasčiausias dvimatės vektorinės regresijos modelis užrašomas tokia lygčių sistema:

$$Y_t = \propto_1 + \ \delta_1 t + \ \sum_{j=1}^{p} \varphi_{1\,j}\, y_{t-j} + \ \sum_{j=1}^{p} \beta_{1\,j}\, x_{t-j} + \ e_{1t}, \tag{4}$$

$$X_t = \propto_2 + \ \delta_2 t + \ \sum_{j=1}^{p} \varphi_{2\,j}\, y_{t-j} + \ \sum_{j=1}^{p} \beta_{2\,j}\, x_{t-j} + \ e_{2t} \tag{5}$$

Šioje dviejų kintamųjų lygčių sistemoje pirma lygtis nurodo tai, jog *X* yra *Y* kitimo priežastis, o antra – atvirkščiai *Y* yra *X* kitimo priežastis. Parametrų φ ir β pirmasis indeksas nurodo, kuriai lygčių sistemos lygčiai jis yra priskiriamas. Atitinkamai yra ir su parametrais α ir δ. Reikia įsidėmėti, kad lagų skaičius *p* yra vienodas visiems kintamiesiems. e_{1t} ir e_{2t} yra atsitiktinės paklaidos, kurios dar yra vadinamos impulsais. Ši lygčių sistema užrašyta tik dviem kintamiesiems, tačiau ji gali būti užrašoma ir didesniam tiriamų kintamųjų skaičiui. Tačiau lygčių skaičius visada bus lygus kintamųjų skaičiui. Norint įvertinti šią sistemą MKM, iš pradžių reikia parinkti *p*. Kadangi parinkus *p,* kointegravimo rangas *r* įprastai yra nežinomas, yra naudinga šiame etape pažymėti tokį VAR pavidalą:

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t, \tag{6}$$

čia $u_t = (u_{1t} \dots, u_{Kt})$ yra nepastebimas nulis, reiškiantis nepriklausomo balto triukšmo procesą su teigiama laiko kovariacijos matrica $E(u_t u_t') = \sum_u$ ir A_i yra (*K* x *K*) matricos koeficientai. Paprastai yra pradedama nuo modelio su tam tikru apibrėžtu maksimaliu lago ilgiu, tarkime p_{max}, ir paeiliui yra tikrinama, kad būtų nustatytas tinkamas modelis, norint dirbti vėlesniuose analizės žingsniuose. Prieš tokios procedūros taikymą, turi būti atliktas p_{max} sprendimas. Kartais šis kiekis yra pasirenkamas pagal tam tikrus teorinius, institucinius argumentus. Pavyzdžiui, vienas iš būdų įtraukti paskutinių metų lagus taip, kad keturi lagai būtų įtraukti ketvirtiniams duomenims ir dvylika lagų gali būti naudojami mėnesiniam modeliui. Parinkus didelį *p*, sumažėja laisvės laipsnių skaičius, o parinkus mažą – padidėja liekamosios paklaidos tikimybė. Jei naudojamas labai didelis p_{max}, gali būti reikalinga ilga bandymų seka, kuri turės įtakos testuojamos sekos bendrai I tipo paklaidai, tai yra, p_{max} pasirinkimas įtakos nepakankamą *p* atrankos tikimybę. Dažniausiai dydžiui *p* nustatyti yra naudojamas Akaikės (*AICC*) kriterijus, Pasirinkus lagų skaičių, vektorinės autoregresijos modelis yra žymimas VAR(*p*).

Regresoriai panaikinami nuosekliai ir tie, kurie sukelia didžiausią duoto kriterijaus sumažėjimą, kol yra galimas kitas sumažinimas. Formaliai:

Žingsnis *j*: panaikinamas x_{kt} , jeigu:

$$CR(1,..,k-1,k+1,..,N-j+1) = \min_{l=1,..,N-j+1} CR(1,..,l-1,l+1,..,N-j+1) \quad (7)$$

Ir

$$CR(1,..,k-1,k+1,..,N-j+1) \leq CR(1,..,N-j+1) \quad (8)$$

Atskiri nuliniai koeficientai taip pat gali būti pasirinkti, remiantis parametro įverčiu *t-statistikomis*. Galima strategija yra nuosekliai panaikinti tuos regresorius su mažiausiomis absoliučiomis t-statistikų reikšmėmis, kol visos *t-statistikos* (absoliučios vertės) yra didesnės negu tam tikra ribinė vertė γ . Reikia pažymėti, kad tik vienetinis regresorius yra eliminuojamas kiekviename žingsnyje. Tada naujos *t-statistikos* yra apskaičiuojamos redukuotam modeliui. Ši strategija yra ekvivalentiška nuosekliam eliminavimui, grindžiamam modelio pasirinkimo kriterijumi, jei ribinė vertė γ yra pasirenkama atitinkamai. Tiksliau, jie rodo, kad pasirinkta $\gamma = \left\{\left[\exp\frac{c_T}{T} - 1\right](T-N+j-1)\right\}^{1/2}$ j-ajame eliminavimo procedūros rezultato žingsnyje tame pačiame galutiniame modelyje, tai taip pat yra gauta pagal nuoseklų pasirinkimo kriterijaus minimizavimą, apibrėžtą pagal sąlygą c_T. Taigi, ribinė reikšmė priklauso nuo atrinkimo kriterijaus cT, imties dydžio ir regresorių skaičiaus modelyje. Ribinė reikšmė *t-statistikoms* atitinka bandymų kritines reikšmes. Šiuose procedūrose gali būti naudojami AICC, HQ ir SC kriterijai.

Į vektorinės autoregresijos modelį įtraukus egzogeninius kintamuosius gauname VAR modelį, kuris yra apibrėžiamas tokia formule:

$$Y_t = a_0 + A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + B_1 X_{t-1} + \cdots B_q X_{t-q} U_t, \quad (9)$$

čia $Y_t \in R^k, X_t \in R^m$ yra egzogeninių kintamųjų vektorius, $a_0 \in R^k$ – atidėjimų koordinačių ašyje vektorius. A_j – k x k koeficiento matricos, B_i yra k x m koeficiento matricos, o $U_t \in R^k$ yra paklaidų vektorius. Modelio korektiškumui lemiama sąlyga yra:

$$E\left[U_t \middle| \{Y_{t-y}\}_{j+1}^{\infty}, \{X_{t-i}\}_{i=1}^{\infty}\right] = 0 (\in R^k) \quad (10)$$

Su tikimybe 1.

Pasikliautinumo intervalai

Norėdami aptikti anomalijas, turime žinoti kokios taškų reikšmės yra normalios, o kokios jau priskiriamos anomalijoms. Tam geriausiai galime panaudoti pasikliautinumo intervalus (angl. *prediction intervals*). Pasikliautinas intervalas – tai intervalas, kuriame tikėtina, yra matuojamo dydžio vidurkis. Tikimybė, jog vidurkis iš tiesų yra šiame intervale vadinama reikšmingumo lygmeniu (dažniausiai yra naudojamas 0,95). Tarkime, kad normaliai pasiskirsčiusių paklaidų intervalai žinant vidurkį ir standartinį nuokrypį gali būti apskaičiuojami tokia formule:

$$\gamma = P(l < X < u) = P\left(\frac{l-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{u-\mu}{\sigma}\right) = P(\frac{l-\mu}{\sigma} < Z < \frac{u-\mu}{\sigma}) \quad (11)$$

Čia $Z = \frac{X-\mu}{\sigma}$, yra standartinis Z rezultatas (angl. *standard score*).

μ – duomenų rinkinio vidurkis;

σ – standartinis nuokrypis;

X – duomenų rinkinio stebinys;

Pagal šią formulę intervalus galima apibūdinti taip:

$$\frac{l-\mu}{\sigma} = -z\,, \frac{u-\mu}{\sigma} = z \quad (12)$$

Z reikšmės, žinant vidurkį ir standartinį duomenų nuokrypį pagal pasikliovimo intervalo dydį pateiktos 2.1. lentelėje.

2.1 lentelė. Z reikšmių ir pasikliovimo intervalų lentelė

Pasikliovimo intervalas	Z reikšmė
75 %	1.15
90 %	1.64
95 %	1.96
99 %	2.58

Tokiu atveju pasikliovimo intervalus galima išreikšti : $[\mu - z\sigma,\ \mu + z\sigma]$.

Būtent pasinaudojant šia formule ir ieškosime anomalijų – taškų, kurie pateks už pasikliautinojo intervalo ribų. Sudarę modelį (VAR ir ARIMA), rasime pritaikytas (angl. *fitted*) duomenų reikšmes, pagal kurias duomenys turėtų būti teisingai pasiskirstę. Palygindami šias reikšmes kartu su jų pasikliautinumo intervalais galima grafiškai rasti, kurie iš duomenų rinkinio taškai nepatenka į nustatytas ribas. Tie taškai bus priskirti išskirtimis.

2.3. Autoregresija su slenkančiu vidurkiu (ARIMA)

Kitas metodas, pagal buvo buvo ieškoma anomalijų - ARIMA – autoregresinio integruoto slenkančio vidurkio (angl. *Autoregressive Integrated Moving Average*) modelio taikymas. Klasikiniai regresijos modeliai dažnai nepakankamai gerai paaiškina laiko eilutės dinamiką. Autoregresiniai modeliai yra paremti idėja, kad dabartinė laiko eilutės reikšmė, gali būti paaiškinta praeities p reikšmėmis x_{t-1}, x_{t-2},...,x_{t-p}.[21] ARIMA modelis laiko eilutę išskaido į autoregresinį procesą AR, aprašantį praeities įvykius, integruotą procesą, padedantį stabilizuoti duomenis, ir slenkančio vidurkio MA procesą, kuris vertina modelio paklaidų įtaką duomenims [22]. Matematinė modelio išraiška yra tokia:

$$\Phi(L)(1-L)^d Y_t = \Theta(L)\varepsilon_t, \tag{13}$$

čia $\Phi(z) = 1 - \varphi_1 z - \cdots - \varphi_p z^p$ - autoregresijos parametrų polinomas;

$\Theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^p$ - slenkančio vidurkio parametrų polinomas;

L – vėlavimo operatorius, su kuriuo $LY_t = Y_{t-1}$;

d žymi integravimo eilę;

ε_t – modelio paklaidų procesą.

Yra daug priežasčių, kodėl ARIMA modelis yra geresnis už kitus modeliavimo būdus laiko eilučių analizėje. Laiko eilučių paklaidoms būdingas koreliuotumas su jų pačių paslinktomis reikšmėmis. Tokia koreliacija neatitinka pagrindinių regresinės analizės prielaidų apie triukšmų nekoreliuotumą. Dėl serijinės koreliacijos regresinė analizė ir standartinė laiko eilučių analizė neefektyvi tarp skirtingų tiesinių operatorių. O kadangi paklaidos gali padėti prognozuoti dabartines paklaidas, šios informacijos naudą galima panaudoti įvedant atitinkamą kintamąjį ARIMA modelio pagalba [23]. Kita serijinės koreliacijos keliama problema yra ta, kad regresinės analizės pagalba

įvertintos standartinės paklaidos nėra vienareikšmės ir korektiškos. Čia taip pat padeda ARIMA modelis.

2.3.1. ARIMA eilės nustatymas

ARIMA modelio eilės nustatymui gali padėti jau anksčiau minėtos autokoreliacijos (ACF) ir dalinės autokoreliacijos (PACF) funkcijos.

AR(p) proceso eilė p nustatoma tiriant dalinės autokoreliacijos koeficientus. AR procesui būdinga tai, kad dalinės autokoreliacijos koeficientas p vėlavimų yra didelis, o likusiuose vėlavimuose dalinė autokoreliacija yra nebereikšminga. MA procesui būdinga tai, kad autokoreliacijos koeficientas yra didelis q vėlavimų, o likusiuose vėlavimuose autokoreliacija yra nebereikšminga.

Taigi, radę ACF ir PACF grafikuose reikšmingumo lygmenį kertančias reikšmes, galima nustatyti, kurios eilės modeliais galėtume aprašyti turimą laiko eilutę.

2.3.2. Eilės parinkimo kriterijai

Paprastai vienai ir tai pačiai laiko eilutei yra keli galimi ARIMA modeliai. Parinkti tinkamiausią ARIMA eilę galima remiantis įvairiais kriterijais, pvz., AIC, BIC [23]. Geriausias modelis tas, kuriam šie rodikliai yra mažiausi. Mūsų atveju – ARIMA modelis buvo parinktas remiantis AIC kriterijumi, kuris yra numatytasis, nepasirinkus kito.

AIC kriterijus

AIC – Akaike informacinis kriterijus yra apibrėžiamas taip:

$$AIC = 2k - 2\ln(L) \tag{14}$$

čia k – modelio parametrų skaičius;

L – maksimizuota modelio tikėtinumo funkcijos reikšmė.

Šis kriterijus įvertina informacijos praradimą taikant modelį realiems duomenims ir gali būti apibūdintas kaip aprašantis kompromisą tarp modelio konstrukcijos poslinkio ir variacijos, arba, kalbant paprasčiau, modelio tikslumo ir sudėtingumo.

2.4. „AnomalyDetection“ paketo metodas

„AnomalyDetection“ R paketas yra skirtas aptikti anomalijas, kurios vertinant iš statistinės pusės, yra sezoniškumo ar tendencijos priežastis. Paketas gali būti naudojamas įvairiose srityse. Jis yra tinkamas aptikti sistemos metrikų anomalijas po naujos programinės įrangos išleidimo, problemoms ekonometrijoje rasti, finansų inžinerijoje, politikos ar socialiniuose mokslų srityse.

Paketo algoritmas yra paremtas sezoniškumo hibrido (angl. *seasonal hybrid ESD*) principu, naudojant bendruosius ESD testus aptikti anomalijas[24]. Šis algoritmas gali aptikti lokalias ir globalias anomalijas. Tai pasiekta panaudojant laiko eilučių dekompoziciją ir naudojant tvirtas statistikos metrikas (pvz., medianą su ESD). Algortimas taip pat yra tinkamas naudoti ilgo laikotarpio duomenims su dideliu dažniu (pvz. šešių mėnesių duomenims kas minutę).

„AnomalyDetection“ paketas gali būti naudojamas ir vektoriaus anomalijoms aptikti. Tai naudinga, kai duomenys neturi laiko žymių. Pakete yra daug rezultatų atvaizdavimo būdų. Naudotojas gali pasirinkti kokios krypties anomalijų ieško, kokio laikotarpio (pvz. paskutinės dienos ar paskutinės valandos).

„AnomalyDetection“ paketo funkcijos

Paketą sudaro dvi pagrindinės funkcijos: *AnomalyDetectionTs()* ir *AnomalyDetectionVec()*. Toliau pateikta tik *AnomalyDetectionTs()* funkcijos panaudojimas ir parametrai, nes šią funkciją naudosime anomalijoms atpažinti.

Naudojimas:

```
AnomalyDetectionTs(x, max_anoms = 0.1, direction = "pos", alpha = 0.05,
  only_last = NULL, threshold = "None", longterm = FALSE, plot = FALSE)
```

Parametrai:

x – dviejų stulpelių laiko eilutė, kurios pirmą stulpelį sudaro laiko žymės, o antrą kintamojo stebiniai;

max_anoms –procentinis dydis, kuris nurodo didžiausią skaičių anomalijų, kuriuos S-H-ESD gali aptikti lyginant su visais duomenimis;

direction – anomalijų kryptis. Galimi pasirinkimai: *‚pos‘* – teigiami svyravimai, *‚neg‘* – neigiami svyravimai, *‚both‘* – abiejų krypčių anomalijų aptikimas;

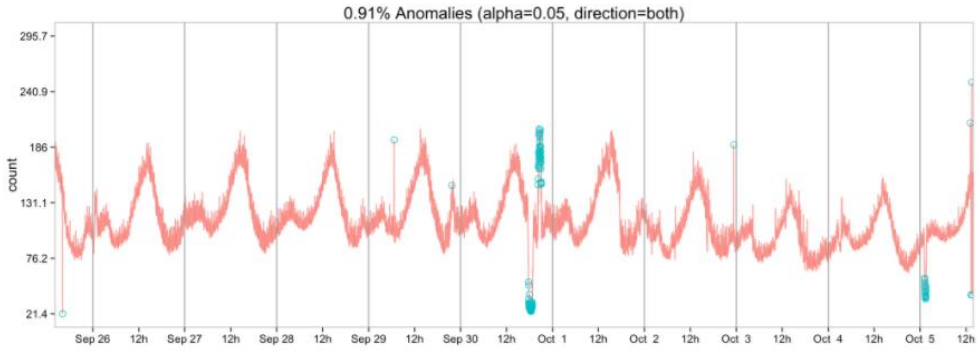
alpha - statistinio reikšmingumo lygis, priimti arba atmesti anomalijas.

only_last – aptikti duomenų anomalijas tik per paskutinę dieną (‚*day*‘) ar valandą (‚*hr*‘) arba visai nenaudoti ‚*None*‘.

threshold – naudojamas tik su teigiama anomalijų kryptimi (‚*pos*‘). Slenkstinės reikšmės pasirinkimas: ‚*None*‘, ‚*med_max*‘, ‚*p95*‘, ‚*p99*‘.

longterm – binominis rodiklis, pasirenkama, jei yra didelis skaičius laiko eilučių;

plot – binominis rodiklis, pasirenkama ar rodyti grafinę analizę su anomalijomis pažymėtomis apskritimais.



2.4 pav. „AnomalyDetection“ paketo grafinės analizės pavyzdys

2.5. Detekcijos metodų kokybės vertinimas – sumaišymo matrica

Sumaišymo matrica (angl. *Confusion matrix*) parodo teisingai ir neteisingai suskirstytų duomenų atvejų skaičių. Dažnai būna, kad sudarytas modelis yra labai tikslus klasifikatorius duomenų rinkiniui, o kitas sudarytas modelis gali tuos pačius duomenis suskirstyti labai prastai. Sumaišymo matricos tikslas – identifikuoti, kokios rūšies klaidos yra būdingos sudarytam modeliui. Ne visos modelių klaidos yra vienodai svarbios. Sumaišymo matrica leidžia pamatyti ir įvertinti klasifikatoriaus prognozavimo klaidas.

Šis vertinimo metodas susideda iš keturių langelių, kurios dažniausiai yra pažymimos angliškais trumpiniais TP (*True Positive*), FP (*False Possitive*), FN (*False Negative*) ir TN (*True negative*). Lentelės pavyzdys pateiktas 2.2. lentelėje.

2.2 lentelė. Sumaišymo matrica

		Prognozuojamos klasės	
		True	**False**
Tikrosios reikšmės	**True**	TP	FN
	False	FP	TN

TP - teisingų atvejų, kurie suklasifikuojami kaip pozityvūs, skaičius;

FP - neteisingų atvejų, kurie suklasifikuojami kaip pozityvūs, skaičius;

FN – neteisingų atvejų, kurie suklasifikuojami kaip negatyvūs, skaičius;

TN – teisingų atvejų, kurie suklasifikuojami kaip negatyvūs, skaičius;

FP ir FN yra modelio daromos klaidos. Priklausomai nuo uždavinio, jos gali turėti skirtingą svarbą. Priklausomai nuo TP, FP, FN, TN naudojimo yra išskiriame tokie modelio įvertinimai:

2.3 lentelė. Sumaišymo matricos įvertinimai.

Pavadinimas	Formulė	Aprašymas
TP – jautrumas	TP/(TP+FN)	Santykis tarp teisingai klasifikuojamų ir visų pozityvių atvejų.
FP – netikras pavojus	FP/(TN + FP)	Santykis tarp klaidingai klasifikuojamų pozityvių ir visų negatyvių atvejų.
FN	FN/(TP+ FN)	Santykis tarp klaidingai klasifikuojamų negatyvių ir visų pozityvių atvejų.
Specifiškumas	TN/(TN + FP)	Santykis tarp klaidingai klasifikuojamų pozityvių ir visų klaidingai klasifikuojamų atvejų.
TN	TN/(TN + FN)	Santykis tarp teisingai klasifikuojamų negatyvių ir visų negatyvių atvejų.
Tikslumas (angl. *precision*)	TP / (TP+FP)	Santykis tarp teisingai klasifikuotų teigiamų ir visų pozityviai klasifikuotų atvejų.
F1	(2 x TP) / (TP + FP) x TP / (TP + FN)) / (TP / (TP + FP) + TP/ (TP+FN))	Įvertinimas, kuris apima ir tikslumą ir pataikymą.
Teisingumas arba prognozavimo teisingumas	(TP + TN) / (TP+FN + TN+FP)	Santykis tarp visų teisingai klasifikuotų ir visų atvejų.
Klaidų dažnis	(FP + FN) / (TP+FN + TN+FP)	Santykis tarp klaidingai klasifikuotų ir visų atvejų.

2.3. lent. pateiktų įvertinimų svarba priklauso nuo sprendžiamo klasifikavimo uždavinio tipo. Pavyzdžiui, blogų įmonių investavimo uždavinyje FP turėtų būti kuo mažesnis, idealu – lygus nuliui. Siekiant rasti anomalijas ir klasifikuoti duomenis, svarbiausi įvertinimai turėtų būti tikslumas ir jautrumas. Šie rodikliai svarbūs dėl didelio duomenų disbalanso. Anomalijos duomenų rinkinyje sudaro tik apie 3% visų duomenų, todėl yra natūralu, kad aukšti įvertinimai tikrinant neigiamas reikšmes bus gaunami nepriklausomai nuo pačio modelio tikslumo. Pavyzdžiui, net jei modelis visiškai neatpažintų anomalijų, jo netikro pavojaus įvertinimas vis tiek bus labai panašus.

3. TYRIMŲ REZULTATAI IR JŲ APTARIMAS

Šioje darbo dalyje bus pateikti atliktų tyrimų rezultatai. Visas tyrimas buvo atliektas naudojantis „RStudio“ programinę įrangą ir jos paketus. Iš pradžių buvo atlikta duomenų analizė, po to praktiškai išbandyti modeliai aprašyti antrajame skyriuje. Kiekvienas modelis buvo įvertintas naudojantis sumaišymo matrica ir jos įvertinimais. Pabaigoje išrinktas modelis, kuris geriausiai atpažino anomalijas.

3.1. Duomenų charakteristika

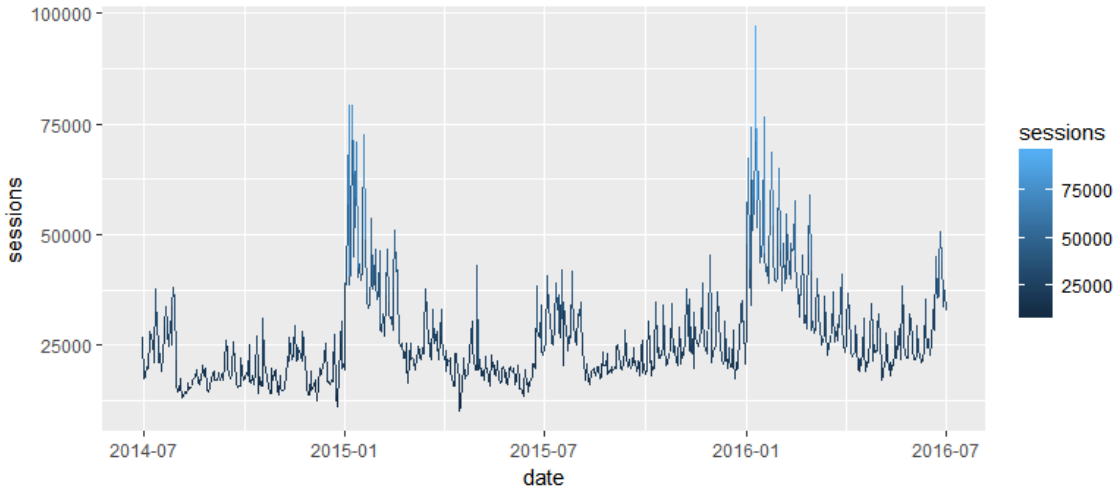
Šiame darbe buvo panaudoti internetinės parduotuvės statistiniai duomenys. Duomenys yra dviejų metų laikotarpio: 2014-06-30 iki 2016-07-01. Iš viso duomenų rinkinį sudaro 9 kintamieji ir 733 eilutės (dieniniai duomenys). Duomenyse yra 26 anomalijos (3.5%) ir 707 normalių duomenų eilutės. Pačioje pradžioje buvo pašalinti 2 kintamieji: *ga.date* ir *reason* . Pirmasis yra data, kuri kartojasi kintamajame *date*, o antrasis yra priklausomo kintamojo paaiškinimas, kuris neturi jokios informacijos naudingos modeliui sudaryti. Galutiniai rinkinio kintamieji:

- *date* – datos formatas.kintamojo data.
- *Sessions* – skaitinė charakteristika. Internetinio puslapio apsilankymų skaičius per dieną;
- *Bounce.rate* – skaitinė charakteristika. Atmetimo rodiklis;
- *Pageviews* – skaitinė charakteristika. Puslapių skaičius, kuris parodo kiek puslapių buvo atidaryta per dieną;
- *ecommerce.conversion.rate* – skaitinė charakteristika. Rodiklis, kuris parodo santykį tarp visų žmonių apsilankiusių puslapyje ir žmonių, kurie ką nors nusipirko;
- *revenue.per.transaction* – skaitinė charakteristika. Rodiklis, kuris parodo vidutinišką pirkimo krepšelio vertę;
- *significant* – dvinaris kintamasis. Reikšmingumas - parodo ar ši eilutė yra anomalija ar ne.

3.1 lentelė. Pateiktos duomenų rinkinio kintamieji su charakteristikomis

	Sessions	Bounce.rate	Pageviews	Ecommerce.conversion.rate	Revenue.per.transaction	significant
Mažiausia reikšmė	9842	9.971	26351	0.169	238.3	0.000
1 ketvirt.	19228	18.051	26351	0.4270	346.6	0.000
Mediana	23245	24.830	116468	0.515	378.4	0.000
Vidurkis	26381	25.783	133788	0.566	381.4	0.035
3 ketvirt.	30363	29.914	156817	0.6545	411.8	0.000
Didžiausia reikšmė	97158	52.125	587730	1.654	563.1	1.000

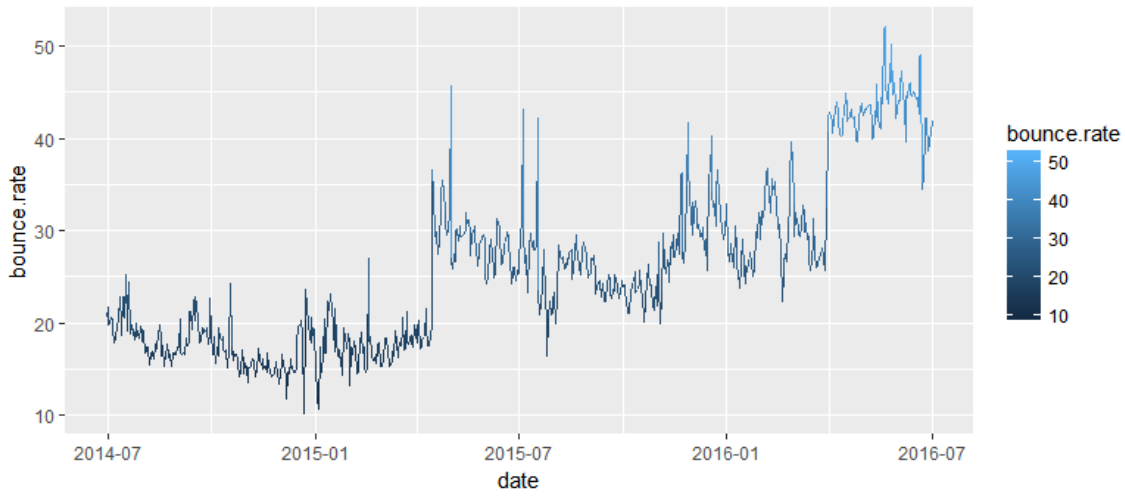
Sesijų skaičiaus (*sessions*) kintamojo mažiausia reikšmė yra lygi 9842, tai reiškia, kad tiek žmonių mažiausiai apsilankė puslapyje per nagrinėjamą laikotarpį. Kadangi mediana yra mažesnė už vidurkį, tai parodo, kad duomenys nėra simetriški ir jiems būdingas kairės pusės asimetrija. Pirmas ir trečias duomenų ketvirčiai parodo, kad visi kintamojo duomenys yra pasiskirstę apie vidurį, nes palyginus – dydis tarp jų nėra labai didelis. Didžiausia reikšmė yra 97 158. Tiek daugiausiai žmonių per vieną parą buvo apsilankę internetinėje parduotuvėje. Visa kintamojo laiko eilutė pateikta 3.1. pav.



3.1 pav. Kintamojo *sessions* laiko eilutė

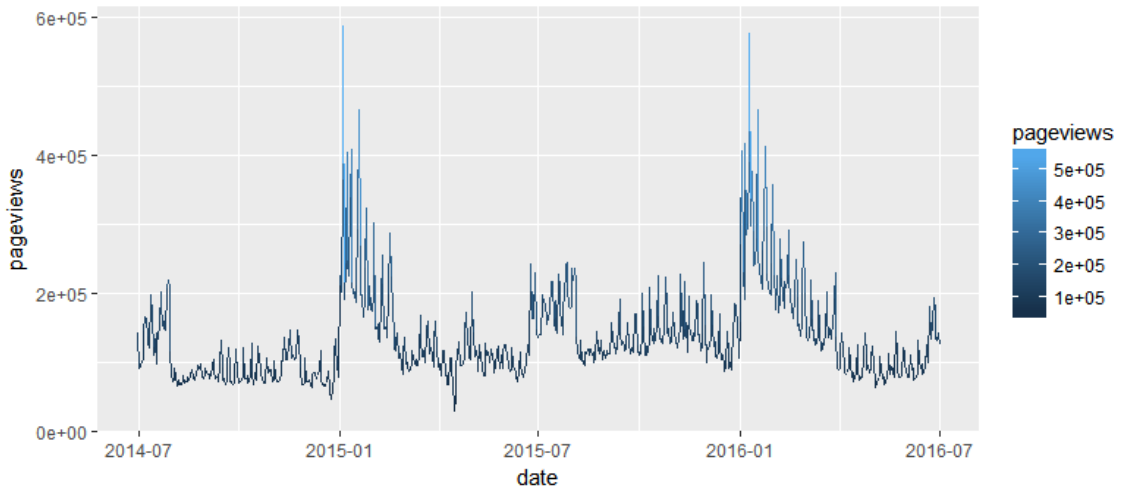
Atmetimo rodiklio (*bounce.rate*) mažiausia reikšmė buvo 9.971, tai reiškia, kad vieną dieną tik beveik 10% visų apsilankiusių tinklalapyje iš karto iš jos išėjo, t. y. net neieškojo prekių, buvo

apsilankęs tik viename internetinės svetainės puslapyje. Ketvirčiai parodo, kad net pusė duomenų yra pasiskirstę gana mažame intervale – [18,05:29,91]. Gana panašūs medianos ir vidurkio rodikliai parodo, kad duomenys yra pasiskirstę beveik simetriškai. Didžiausia reikšmė parodo, kad vieną dieną, daugiau negu pusė žmonių, kurie apsilankė puslapyje (52,125%) iš karto iš jo išėjo. Gali būti, kad tai yra netinkamos reklamos rezultatas. Tai parodo, kad žmonės, kurie pateko į svetainę tikriausiai net neieškojo prekių, kokias siūlo ši parduotuvė, o gal puslapio struktūra ar dizainas nepatiko lankytojams. Visa kintamojo laiko eilutė pateikta 3.2. pav.



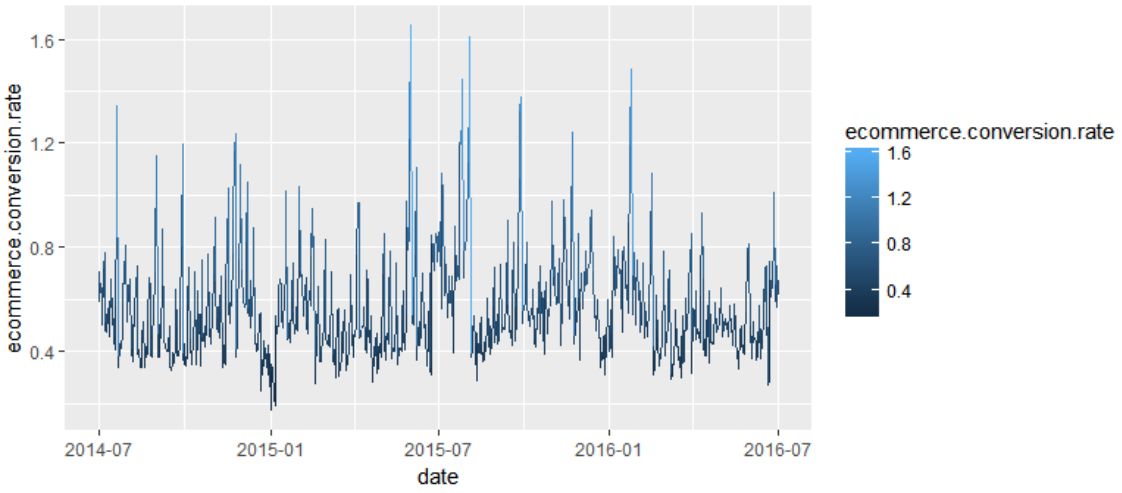
3.2 pav. Kintamojo *bounce.rate* laiko eilutė

Aplankytų puslapių skaičius (*pageviews*) charakteristika parodo, kad mažiausiai 26 351 puslapių buvo atidaryta per parą. Pirmas ir trečias ketvirtis parodo, kad didelė dalis duomenų yra pasiskirsčiusi mažame intervale, tarp jų skirtumas tik apie 70 000. Tai yra gana mažas intervalas palyginti skirtumą tarp didžiausios ir mažiausios reikšmės ar medianos ir vidurkio, kurie taip pat, palyginus su didžiausia reikšme yra maži. Visa kintamojo laiko eilutė pateikta 3.3. pav.



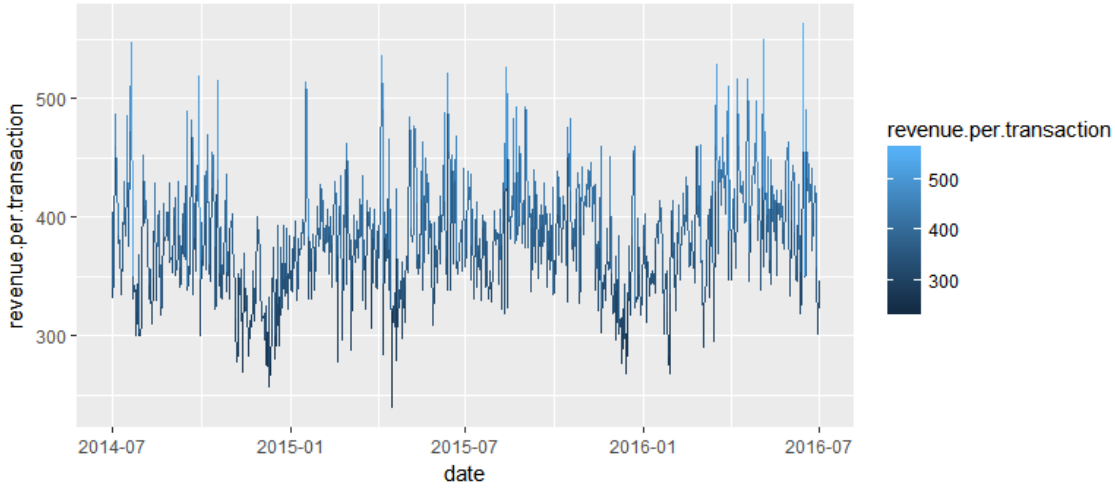
3.3 pav. Kintamojo *pageviews* laiko eilutė

Konversijos rodiklis (*ecommerce.conversion.rate*) mažiausia reikšmė parodo, kad tik 17% iš visų vieną dieną apsilankiusių pirko prekes. Tačiau pagal pirmo ketvirčio rodiklį žinoma, kad mažas konversijos rodiklis yra gana retas reiškinys šiai internetinei parduotuvei. Gana didelis vidurkis ir mediana parodo, kad dažniausiai bent pusė aplankiusių internetinę svetainę lankytojų joje kažką nusiperka. Visa kintamojo laiko eilutė pateikta 3.4. pav.



3.4 pav. Kintamojo *ecommerce.conversion.rate* laiko eilutė

Vidutinė krepšelio vertė (*revenue.per.transaction*) statistika parodo, kad šios įmonės pirkėjai išleidžia didelius pinigus. Deja, nėra žinoma kokia valiuta šis rodiklis yra pateiktas. Tačiau, jei tai Europos Sąjungos šalyje esanti bendrovė, tai tikriausiai yra eurai. Didelė suma parodo, kad arba parduotuvėje yra pardavinėjamos aukštos kainos produkcija, arba pirkėjai perka didelius kiekius produktų. Labai mažas skirtumas tarp medianos ir vidurkio parodo, kad duomenys yra pasiskirstę simetriškai. Visa kintamojo laiko eilutė pateikta 3.5. pav.



3.5 pav. Kintamojo *revenue.per.transaction* laiko eilutė

Reikšmingumo (*significant*) rodiklio statistika yra mažai naudinga, tačiau galima pamatyti, kad tik 3,55% duomenų yra anomalijų. Tai parodo, kad duomenyse yra didelis disbalansas.

Duomenų stacionarumo tikrinimas

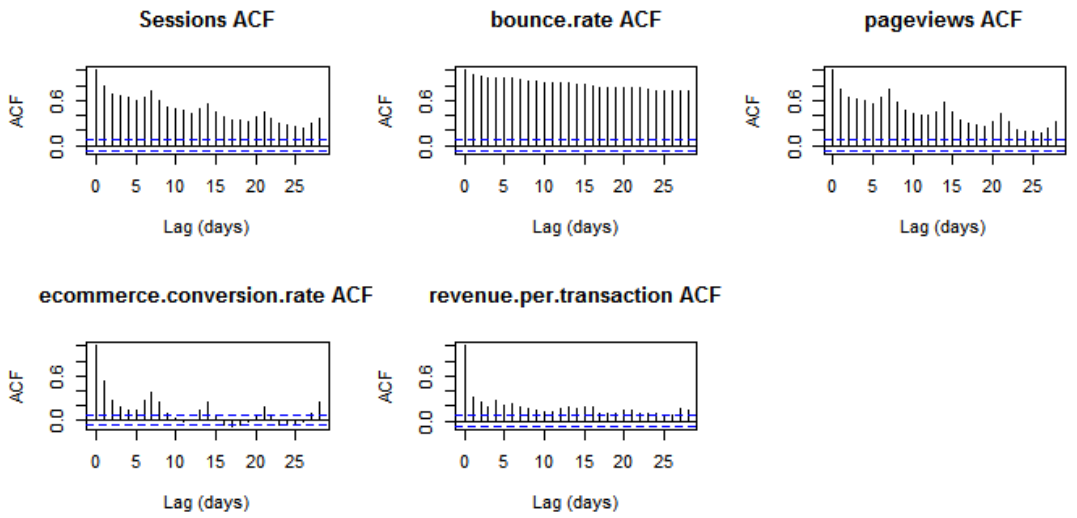
Norint sudaryti modelius reikia sutvarkyti duomenis. Dažniausiai tik sugeneruoti duomenys gali būti iškarto tinkami modeliams sudaryti. Pagrindiniai keli uždaviniai paruošiant duomenis yra:

- Patikrinti kintamųjų tipus (Pakeisti į skaitinius, jei yra tekstinių, simbolinių duomenų);
- Patikrinti duomenų stacionarumą;

Tik atlikus visus uždavinius galima sudaryti modelius. Uždaviniai gali kisti priklausomai nuo duomenų tipo, metodo, kuriuo modeliuojami duomenys.

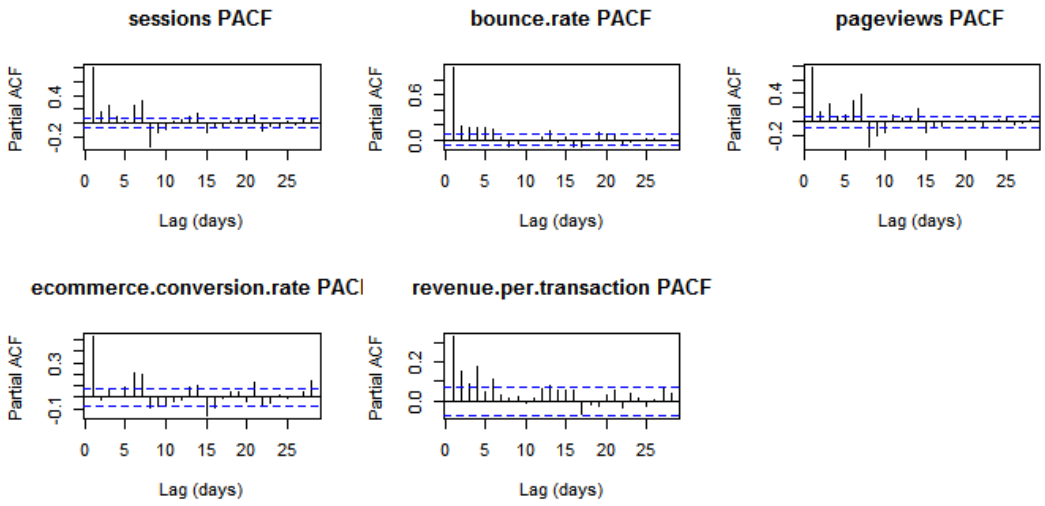
Šiame duomenų rinkinyje yra dviejų tipų kintamieji – datos ir skaitiniai. Šie tipai yra tinkami sudaryti modeliui, todėl nieko keisti nereikia.

Toliau buvo patikrintas duomenų stacionarumas. Iš duomenų grafikų , anksčiau, galima pastebėti, kad duomenys nėra stacionarūs, tačiau norint tuo įsitikinti buvo panaudoti ACF ir PACF grafinė analizė (3.6. pav.)



3.6 pav. Kintamųjų autokoreliacijos grafinė analizė

Atlikus autokoreliacijos grafinę analizė įsitikinta, kad visi kintamieji yra nestacionarūs. Tai patvirtina ACF reikšmės, kurios visos viršija slenkstinę ribą. Taip pat buvo atlikta dalinė autokoreliacijos(PACF) grafinė analizė (3.7. pav.)



3.7 pav. Kintamųjų dalinės autokoreliacijos grafinė analizė

PACF grafinė analizė taip pat parodo duomenų nestacionarumą. Dauguma duomenų neperžengia slenkstinės reikšmės. Tačiau vien šių grafinių analizių neužtenka duomenų stacionarumui patikrinti. Todėl norėdami įsitikinti ar duomenys yra tikrai nestacionarūs buvo atlikti specialūs tam testai: *Box test* ir *adf test*. Šių testų rezultatai pateikti 3.8. – 3.9. pav.

```
        Box-Ljung test

data:  sessions
X-squared = 5471.3, df = 50, p-value < 2.2e-16

        Box-Ljung test

data:  bounce.rate
X-squared = 21019, df = 50, p-value < 2.2e-16

        Box-Ljung test

data:  pageviews
X-squared = 4577.5, df = 50, p-value < 2.2e-16
```

```
        Augmented Dickey-Fuller Test

data:  sessions
Dickey-Fuller = -4.4264, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test

data:  bounce.rate
Dickey-Fuller = -3.6883, Lag order = 9, p-value = 0.02458
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test

data:  pageviews
Dickey-Fuller = -4.8019, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

3.8 pav. Kintamųjų *sessions, bounce.rate ir pageviews* stacionarumo testai

```
        Box-Ljung test

data:  ecommerce.conversion.rate
X-squared = 867.92, df = 50, p-value < 2.2e-16
```

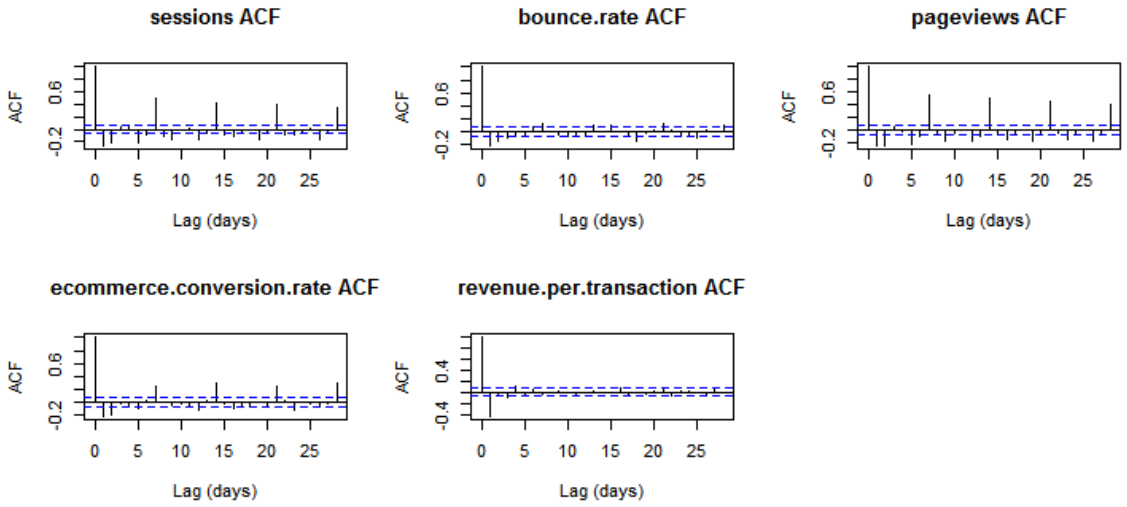
```
        Augmented Dickey-Fuller Test

data:  ecommerce.conversion.rate
Dickey-Fuller = -6.1661, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

```
        Box-Ljung test

data:  revenue.per.transaction
X-squared = 655.53, df = 50, p-value < 2.2e-16
```

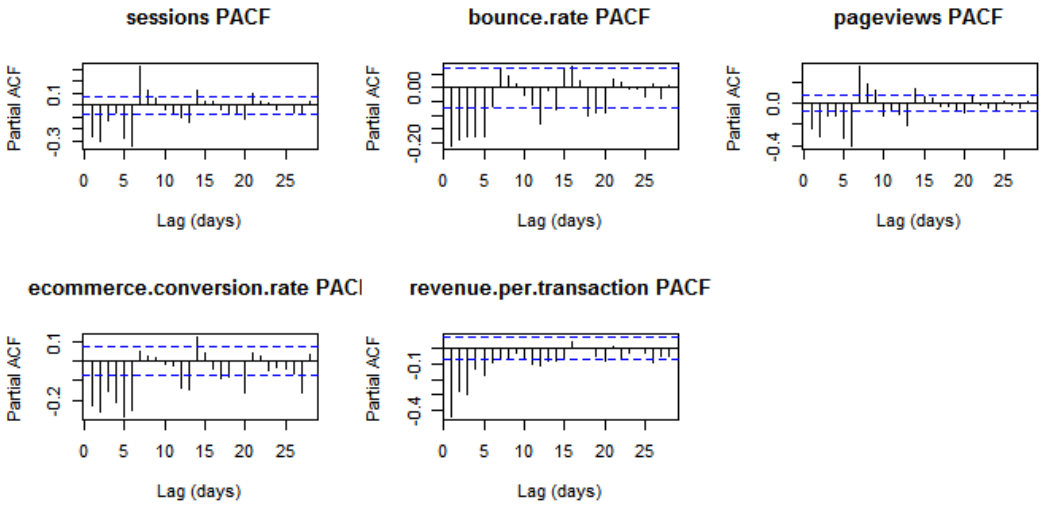
```
        Augmented Dickey-Fuller Test

data:  revenue.per.transaction
Dickey-Fuller = -5.8332, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

3.9 pav. Kintamųjų *ecommerce.conversion.rate* ir *revenue.per.transaction* stacionarumo testai

Abu testai rodo, kad visos laiko eilutės yra stacionarios (p-value < 0,05). Tačiau panaudojant nediferencijuotus duomenis, buvo gauti labai prasti rezultatai. Dauguma modelių atpažino tik po kelias anomalijas, taip pat ARIMA modelis parenka *d=1* parametrą, kas parodo, kad modelis atpažįsta juos kaip nestacionarius. Tam, kad ARIMA modeliui nereikėtų diferencijuoti duomenų ir dėl ACF ir PACF grafikų duomenis vieną kartą diferencijuosime. Po šio proceso dar kartą yra atliekama ACF ir PACF grafinė analizė. (3.10. pav. ir 3.11. pav.)



3.10 pav. Kintamųjų autokoreliacijos grafinė analizė po diferencijavimo



3.11 pav. Kintamųjų dalinės autokoreliacijos grafinė analizė po diferencijavimo

Stacionarumo testų rezultatai po diferencijavimo nepasikeitė, duomenys pagal testus yra stacionarūs.(3.12. pav. ir 3.13. pav.)

```
        Box-Ljung test

data:  sessions
X-squared = 1109.6, df = 50, p-value < 2.2e-16

        Box-Ljung test

data:  bounce.rate
X-squared = 220.34, df = 50, p-value < 2.2e-16

        Box-Ljung test

data:  pageviews
X-squared = 1333.5, df = 50, p-value < 2.2e-16

        Box-Ljung test

data:  ecommerce.conversion.rate
X-squared = 612.26, df = 50, p-value < 2.2e-16

        Box-Ljung test

data:  revenue.per.transaction
X-squared = 222.27, df = 50, p-value < 2.2e-16
```

3.12 pav. Ljung-Box stacionarumo testo rezultatai po diferencijavimo

```
        Augmented Dickey-Fuller Test

data:  sessions
Dickey-Fuller = -8.7996, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test

data:  bounce.rate
Dickey-Fuller = -9.8634, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test

data:  pageviews
Dickey-Fuller = -8.5945, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test

data:  ecommerce.conversion.rate
Dickey-Fuller = -11.024, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test

data:  revenue.per.transaction
Dickey-Fuller = -12.389, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

3.13 pav. Augmented Dickey-Fuller stacionarumo testo rezultatai po diferencijavimo

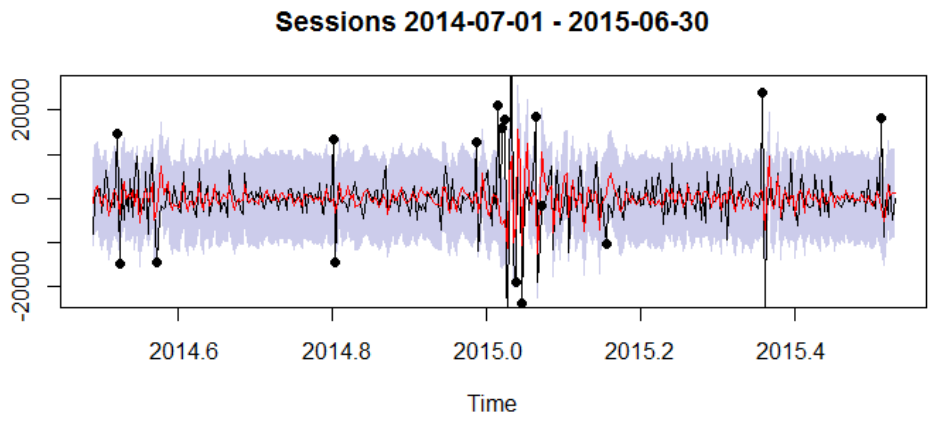
Atlikus diferencijavimą, rezultatai rodo, kad duomenys tikrai yra stacionarūs. Nors tarp jų ir yra reikšmingų lagų, kurie viršija slenkstinę ribą, tačiau jie greitai „užgęsta“. Iš ACF ir PACF testų sužinota, kad duomenims yra būdingas savaitinis periodiškumas.

Atlikus visus žingsnius reikalingus duomenų paruošimui modeliuoti, galima sudaryti modelius.

3.2. Anomalijų aptikimo rezultatai naudojant ARIMA modelį ir pasikliovimo intervalus

Pirmasis metodas aptikti anomalijoms duomenyse yra autoregresijos modelis su slenkančiu vidurkiu. Iš pradžių buvo sudaryti kiekvieno kintamojo modeliai, su jų laiko eilutėmis ir vienu išoriniu regresoriumi – savaitės diena (*day*). Anomalijų aptikimui bus naudojami pasikliovimo intervalai. Anomalijomis bus priskirti duomenų taškai, kurie bus už modelio pritaikytų taškų(angl. *fitted*) pasikliovimo intervalo ribų. Norint sudaryti tikslesnius modelius duomenims, jie buvo padalinti į metinius: 2 duomenų rinkinius. Modelių parametrai buvo nustatyti naudojantis *auto.arima()* funkcija. Bus naudojamas standartinis 90 %, 95%, 99 % pasikliovimo intervalai anomalijoms aptikti. Toliau bus pateikti tik 95 % kiekvieno kintamojo rezultatai ir bendri rezultatai panaudojant visus pasikliovimo intervalus. Modeliams įvertinti naudojama sumaišymo matrica su dviem įvertinimais – specifiškumu ir tikslumu.

Pirmojo kintamojo ARIMA modelių parametrai atlikus funkciją: 1 metų – ARIMA(2, 0, 0) $(1,0,1)_7$. Tai parodo, kad bus naudojamas antros eilės autoregresijos lagas ir duomenims yra būdingas savaitinis sezoniškumas.



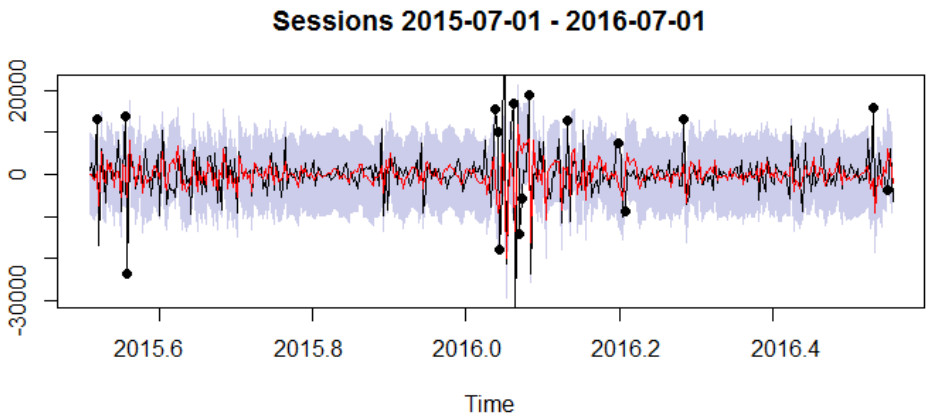
3.14 pav. Kintamojo *sessions* 2014-07-01 – 2015-06-30 grafiniai rezultatai

3.14. pav. matyti kurie taškai yra už pasikliautinumo ribų ir buvo priskirti anomalijoms (pažymėti juodais taškais.) 3.2. lent. pateikta sumaišymo matrica. Buvo aptikta 19 anomalijų, tačiau tik 8 iš jų atitiko anomalijas palyginus su tikrais duomenimis.

3.2 lentelė. Kintamojo *sessions* pirmų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	8	5
	False	11	341
Įvertinimai			
Tikslumas (angl. *precision*)		61.2%	
Specifiškumas		68.8%	

Antrų metų kintamojo ARIMA modelių parametrai atlikus funkciją: ARIMA(0,0,1)(2,0,1)$_7$. Tai parodo, kad bus naudojamas pirmos eilės slenkančio vidurkio lagas ir duomenims yra būdingas savaitinis sezoniškumas.



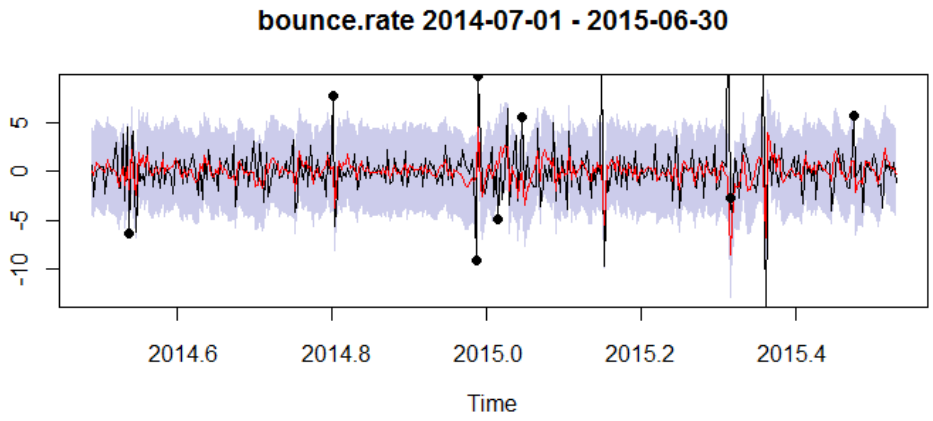
3.15 pav. Kintamojo *sessions* 2015-07-01 – 2016-07-01 grafiniai rezultatai

Per antruosius nagrinėjamus metus anomalijas aptikti buvo dar sunkiau. Nors bendrai buvo aptikta panašus skaičius anomalijų (18), tačiau tik 2 iš jų atitiko anomalijas tikruosiuose duomenyse. Toks santykis parodo labai mažą modelio tikslumą. Modelio specifiškumas yra didelis palyginti su jo tikslumu.

3.3 lentelė. Kintamojo *sessions* antrų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	2	11
	False	16	338
Įvertinimai			
Tikslumas (angl. *precision*)		15.4%	
Specifiškumas		59.3%	

Pirmų metų kintamojo *bounce.rate* ARIMA modelių parametrai atlikus funkciją: ARIMA $(1,0,1)(0,0,1)_7$. Tai parodo, kad bus naudojamas pirmos eilės slenkančio vidurkio lagas ir pirmos eilės slenkančio vidurkio lagas. Duomenims yra būdingas savaitinis sezoniškumas.



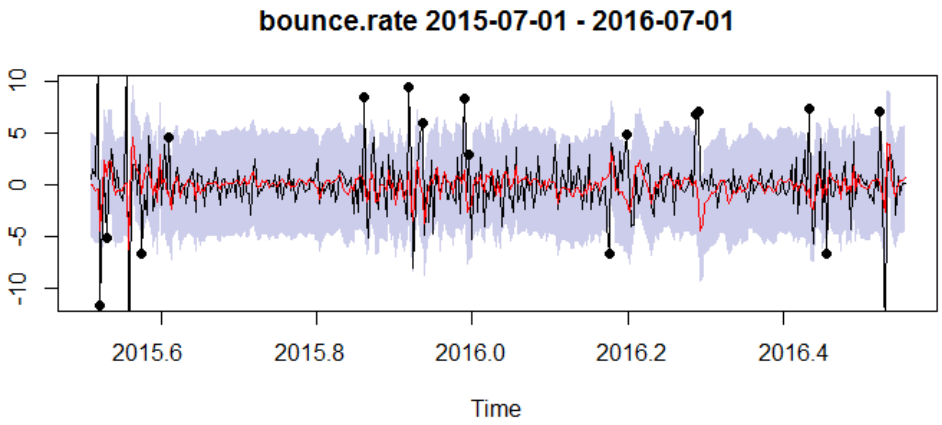
3.16 pav. Kintamojo *bounce.rate* 2014-07-01 – 2015-06-30 grafiniai rezultatai

Atlikta analizė parodė, kad *bounce.rate* kintamojo pagalba aptikti anomalijas yra sunkiau negu sesijų skaičius. Per pirmus metus buvo rasta 12 anomalijų, iš kurių tik 3 atitiko anomalijas tikruose duomenyse. Šio modelio tikslumas taip pat yra mažas.

3.4 lentelė. Kintamojo *bounce.rate* pirmų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	3	10
	False	9	343
Įvertinimai			
Tikslumas (angl. *precision*)		23.1%	
Specifiškumas		47.4%	

Antrų metų kintamojo *bounce.rate* ARIMA modelių parametrai atlikus funkciją: ARIMA $(0,0,2)(2,0,0)_7$. Tai parodo, kad bus naudojamas antros eilės slenkančio vidurkio lagas. Duomenims yra būdingas savaitinis sezoniškumas.



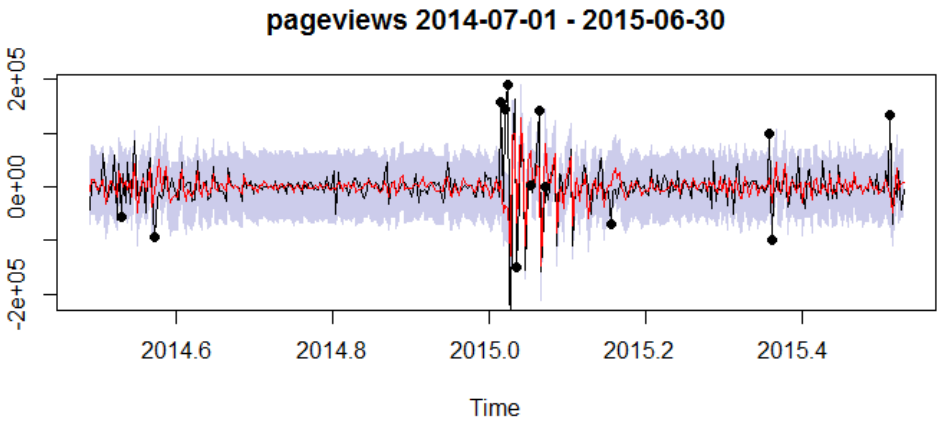
3.17 pav. Kintamojo *bounce.rate* 2015-07-01 – 2016-07-01 grafiniai rezultatai

Per antruosius metus kintamojo *bounce.rate* laiko eilutėje buvo rasta daugiau anomalijų – 20. Patikrinus su tikraisiais duomenimis 5 iš jų buvo teisingos. Didelis modelio specifiškumas parodo, kad nors modelis ir randa daug anomalijų, tačiau tik mažas skaičius jų yra tikrų.

3.5 lentelė. Kintamojo *bounce.rate* antrų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	5	8
	False	15	339
Įvertinimai			
Tikslumas (angl. *precision*)		38.5%	
Specifiškumas		65.2%	

Pirmų metų kintamojo *pageviews* ARIMA modelių parametrai atlikus funkciją: ARIMA $(2,0,1)(2,0,2)_7$. Tai parodo, kad bus naudojamas antros eilės autoregresijos lagas ir pirmos eilės slenkančio vidurkio lagas. Duomenims yra būdingas savaitinis sezoniškumas.



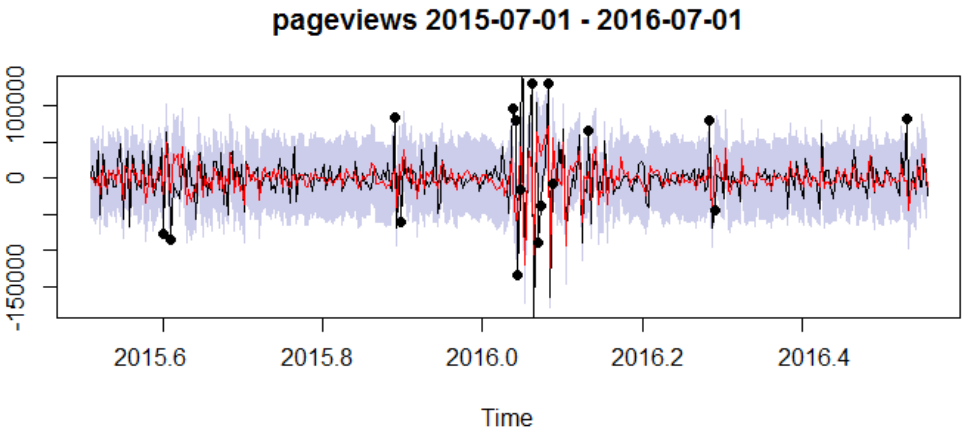
3.18 pav. Kintamojo *pageviews* 2014-07-01 – 2015-06-30 grafiniai rezultatai

Per pirmuosius metus kintamojo *pageviews* laiko eilutėje buvo 14 anomalijų. Net pusė iš jų patikrinus su tikraisiais duomenimis atitiko anomalijas. Modelis palyginus su ankstesniais yra gana tikslus ir specifiškas.

3.6 lentelė. Kintamojo *pageviews* pirmų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	7	6
	False	7	345
Įvertinimai			
Tikslumas (angl. *precision*)		53.8%	
Specifiškumas		53.8%	

Antrų metų kintamojo *pageviews* ARIMA modelių parametrai atlikus funkciją: ARIMA $(0,0,1)(1,0,1)_7$. Tai parodo, kad bus naudojamas pirmos eilės slenkančio vidurkio lagas. Duomenims yra būdingas savaitinis sezoniškumas.



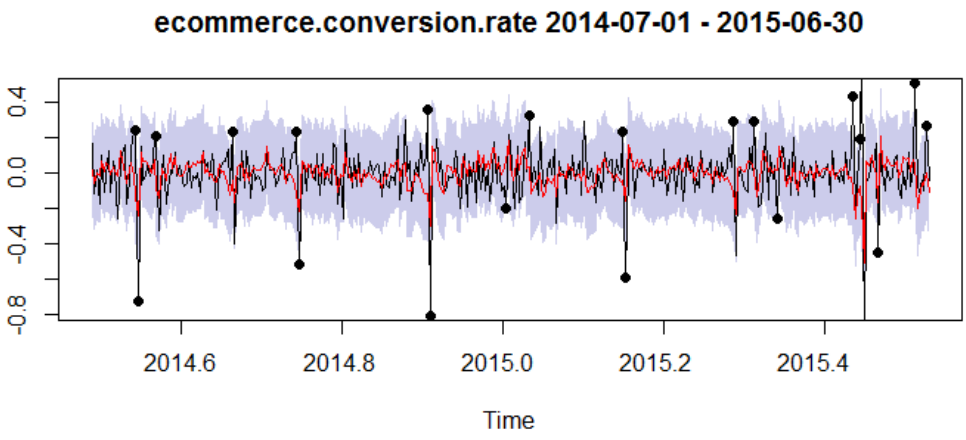
3.19 pav. Kintamojo *pageviews* 2015-07-01 – 2016-07-01 grafiniai rezultatai

Per antruosius metus kintamojo *pageviews* laiko eilutėje buvo aptikta 19 anomalijų. Tačiau priešingai nei per pirmuosius metus, šį kartą tik 2 iš jų buvo anomalijos iš tikrųjų. Tai lėmė mažą modelio tikslumą ir didelį specifiškumą.

3.7 lentelė. Kintamojo *pageviews* antrų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	2	11
	False	17	337
Įvertinimai			
Tikslumas (angl. *precision*)		15.4%	
Specifiškumas		60.7%	

Pirmųjų metų kintamojo *ecommerce.conversion.rate* ARIMA modelių parametrai atlikus funkciją: ARIMA (1,0,1)(1,0,1) $_7$. Tai parodo, kad bus naudojamas pirmos eilės autoregresijos lagas ir pirmos eilės slenkančio vidurkio lagas. Duomenims yra būdingas savaitinis sezoniškumas.



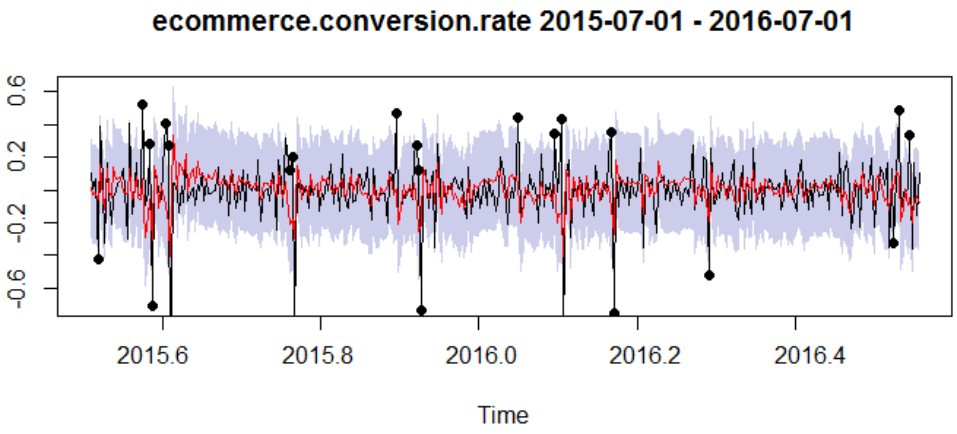
3.20 pav. Kintamojo *ecommerce.conversion.rate* 2014-07-01 – 2015-06-30 grafiniai rezultatai

Per pirmuosius nagrinėjamus metus laiko eilutėje buvo atpažinta 22 anomalijos. Patikrinus anomalijas su tikraisiais duomenimis paaiškėjo, kad tik 3 iš jų yra anomalijos.

3.8 lentelė. Kintamojo *ecommerce.conversion.rate* pirmų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	3	10
	False	19	333
Įvertinimai			
Tikslumas (angl. *precision*)		23.1%	
Specifiškumas		65.5%	

Antrųjų metų kintamojo *ecommerce.conversion.rate* ARIMA modelių parametrai atlikus funkciją: ARIMA (3,0,4). Tai parodo, kad bus naudojamas trečios eilės autoregresijos lagas ir ketvirtos eilės slenkančio vidurkio lagas.



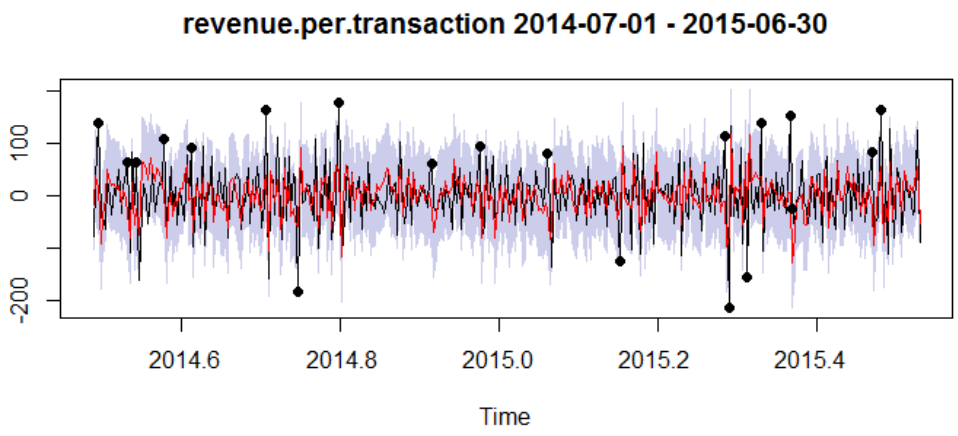
3.21 pav. Kintamojo *ecommerce.conversion.rate* 2015-07-01 – 2016-07-01 grafiniai rezultatai

Per antruosius metus kintamojo *ecommerce.conversion.rate* laiko eilutėje buvo aptikta 24 anomalijos. Tačiau patikrinus su realiais duomenimis, anomalijos buvo tik 3 taškai. Šiam modeliui būdingos pirmo tipo klaidos, kas lėmė didelį specifiškumą.

3.9 lentelė. Kintamojo *ecommerce.conversion.rate* antrų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	3	10
	False	21	333
Įvertinimai			
Tikslumas (angl. *precision*)		23.1%	
Specifiškumas		67.7%	

Paskutinio kintamojo pirmųjų metų kintamojo ARIMA modelių parametrai atlikus funkciją: ARIMA $(0,0,1)(1,0,0)_7$. Tai parodo, kad bus naudojamas pirmos eilės slenkančio vidurkio lagas. Duomenims yra būdingas savaitinis sezoniškumas.



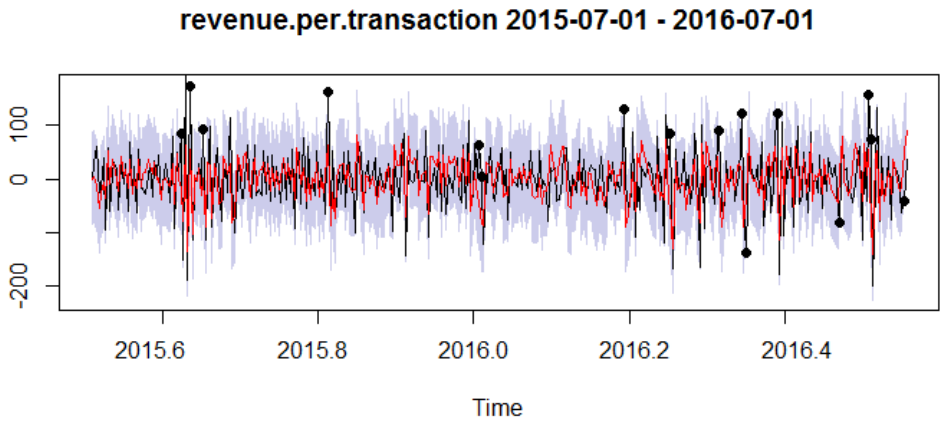
3.22 pav. Kintamojo *revenue.per.transaction* 2014-07-01 – 2015-06-30 grafiniai rezultatai

Per pirmuosius metus kintamojo *revenue.per.transaction* laiko eilutėje buvo rastos 20 anomalijos. Patikrinus su realiais duomenimis, paaiškėjo, kad tik 1 iš jų atitiko (7.7 % tikslumas).

3.10 lentelė. Kintamojo *revenue.per.transaction* pirmų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	1	12
	False	19	333
Įvertinimai			
Tikslumas (angl. *precision*)		7.7%	
Specifiškumas		61.3%	

Paskutinio kintamojo pirmųjų metų kintamojo ARIMA modelių parametrai atlikus funkciją: ARIMA $(0,0,1)(0,0,2)_7$. Tai parodo, kad bus naudojamas pirmos eilės slenkančio vidurkio lagas. Duomenims yra būdingas savaitinis sezoniškumas.



3.23 pav. Kintamojo *revenue.per.transaction* 2015-07-01 – 2016-07-01 grafiniai rezultatai

Antruosiuose metuose kintamojo *revenue.per.transaction* laiko eilutėje rasta 17 anomalijų, tačiau nei viena iš jų nesutapo su anomalijomis tikruose duomenyse. Modeliui būdingas didelis pirmo tipo klaidų skaičius, ką parodo specifiškumas.

3.11 lentelė. Kintamojo *revenue.per.transaction* sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	0	13
	False	17	337
Įvertinimai			
Tikslumas (angl. *precision*)		0%	
Specifiškumas		56.7%	

Atlikus visų kintamųjų anomalijų aptikimą naudojantis ARIMA ir pasikliautinuoju intervalu yra bendrai įvertinama kiek iš viso anomalijų aptiko šis metodas ir kiek iš jų sutapo su tikriausiai duomenimis. Tam sužinoti buvo sudėtos visos anomalijos ir pašalintos pasikartojančios. Visų kintamųjų sumaišymo matricos rezultatai pateikti 16 lent.. Nors iš viso buvo aptikta net 134 anomalijos, tačiau tik 16 iš jų sutapo su anomalijomis tikruose duomenyse (92.2 % specifiškumas).

Palyginimui buvo pakeista pasikliautinas intervalas į 90 % ir 99 % (3.13 ir 3.14 lent.). Tai leido padėti atrinkti, kuris iš pasikliautinųmų intervalų būtų tinkamiausias atpažinti anomalijas, kad santykis tarp atpažintų teisingai ir klaidingai būtų geriausias. Padidinus intervalą iki 99 %, buvo aptikta 14 anomalijų, tačiau neteisingai atpažintų anomalijų skaičius lengviau sumažėjo (iki 54). Sumažinus

intervalą iki 90 % anomalijų buvo atrasta tik 2 daugiau, tačiau neteisingai identifikuotų anomalijų skaičius padidėjo net 170.

3.12 lentelė. Visų kintamųjų sumaišymo matricos, jautrumo ir tikslumo įvertinimų bendri rezultatai su 95 % pasikliautinuoju intervalu

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	16	10
	False	118	588
Įvertinimai			
Tikslumas (angl. *precision*)		61.5%	
Specifiškumas		92.2%	

3.13 lentelė. Visų kintamųjų sumaišymo matricos, jautrumo ir tikslumo įvertinimų bendri rezultatai su 90 % pasikliautinuoju intervalu

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	18	8
	False	170	536
Įvertinimai			
Tikslumas (angl. *precision*)		69.2%	
Specifiškumas		95.5%	

3.14 lentelė. Visų kintamųjų sumaišymo matricos, jautrumo ir tikslumo įvertinimų bendri rezultatai su 99 % pasikliautinuoju intervalu

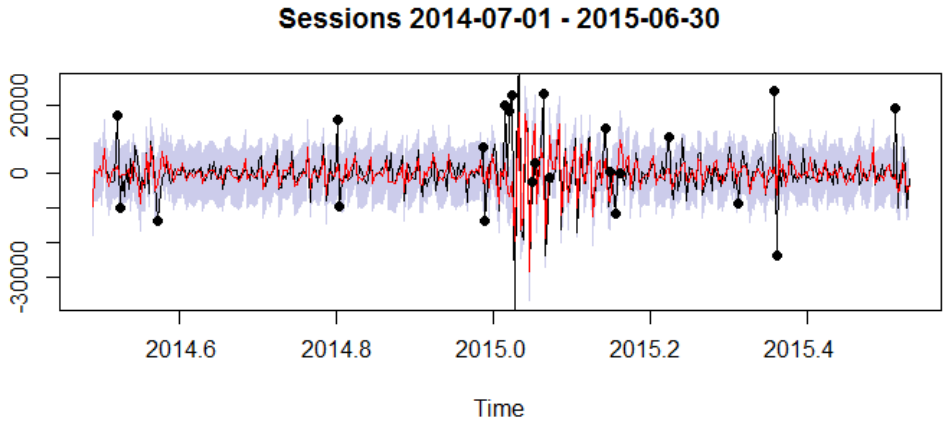
		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	14	12
	False	54	588
Įvertinimai			
Tikslumas (angl. *precision*)		53.8%	
Specifiškumas		81.8%	

3.3. Anomalijų aptikimo rezultatai naudojant VAR ir pasikliovimo intervalus

Šioje darbo dalyje bus naudojamas vektorinės autoregresijos modelis su išoriniais kintamaisiais (VAR). Kaip ir ARIMA modelyje, bus naudojamas tas pats išorinis kintamasis – savaitės diena (*day*). Duomenys bus padalinti taip pat kaip ARIMA modelyje – po 1 metus. Bus naudojami tie patys pasikliovimo intervalai – 90 %, 95% ir 99%. Toliau bus pateikti tik 95 % pasikliovimo interval atskirų kintamųjų rezultatai, o pabaigoje bendri visų intervalų rezultatai.

3.3.1. Kintamųjų rezultatai pirmaisiais metais (2014-07-01 – 2015-06-30)

Atlikus automatinę VAR modelio parametro nustatymo funkciją, buvo gauta, kad kintamiesiems pirmųjų metų duomenis bus naudojamas modelis VAR(7). Tai reiškia, kad bus naudojama 7 dienų lago vektorinė autoregresijos analizė.

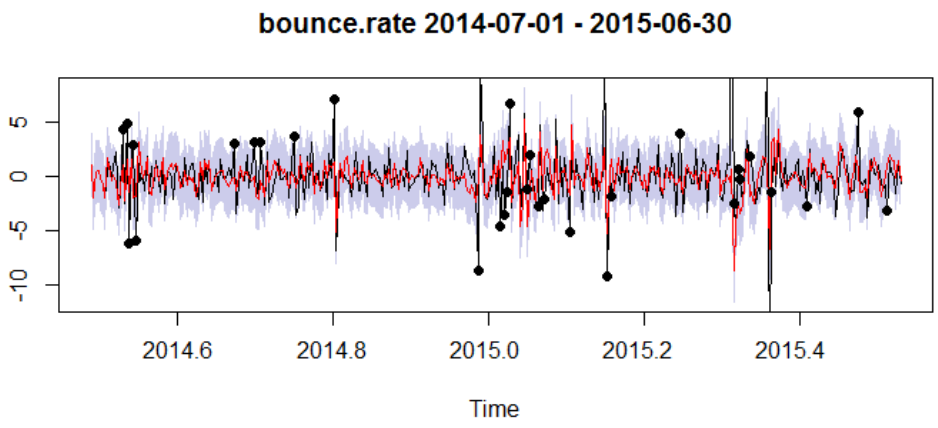


3.24 pav. Kintamojo *sessions* 2014-07-01 – 2015-06-30 grafiniai rezultatai

Pirmojo kintamojo 2014-07-01 – 2015-06-30 laikotarpiu buvo aptiktos 25 anomalijos. Palyginus šias anomalijas su tikraisiais duomenimis buvo rasti 9 sutapimai. Tai reiškia, kad remiantis šiuo kriterijumi buvo atpažintos 9 anomalijos. Modelis yra gana tikslus (69.2% tikslumas). Tačiau modeliui yra būdingos pirmo tipo klaidos – aukštas specifiškumas (80 %).

3.15 lentelė. Kintamojo *sessions* pirmų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	9	4
	False	16	336
Įvertinimai			
Tikslumas (angl. *precision*)		69.2%	
Specifiškumas		80%	

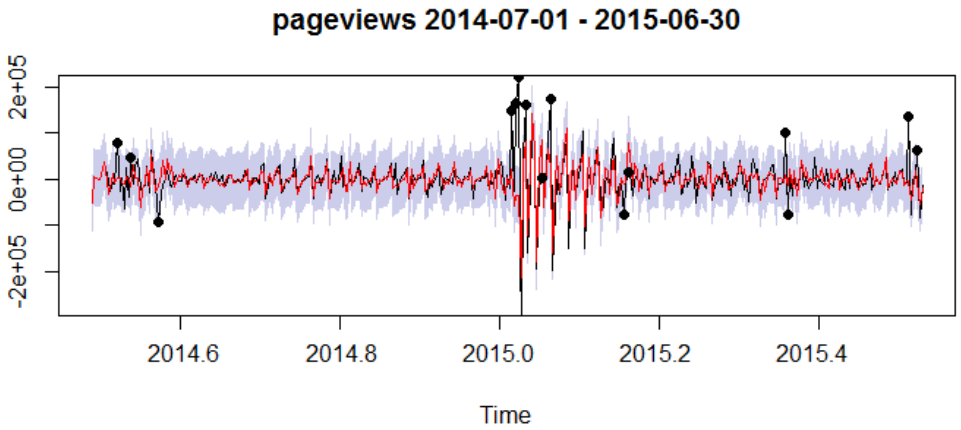


3.25 pav. Kintamojo *bounce.rate* 2014-07-01 – 2015-06-30 grafiniai rezultatai

Atlikus antrojo kintamojo *bounce.rate* VAR modelio analizę, buvo aptiktos net 36 anomalijos. Palyginus su tikrosiomis reikšmėmis buvo įsitikinta, kad buvo atrastos 7 anomalijos. Šio modelio tikslumas yra mažesnis, tačiau aukštas specifiškumas.

3.16 lentelė. Kintamojo *bounce.rate* pirmų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	7	6
	False	29	323
Įvertinimai			
Tikslumas (angl. *precision*)		53.8%	
Specifiškumas		82.9%	

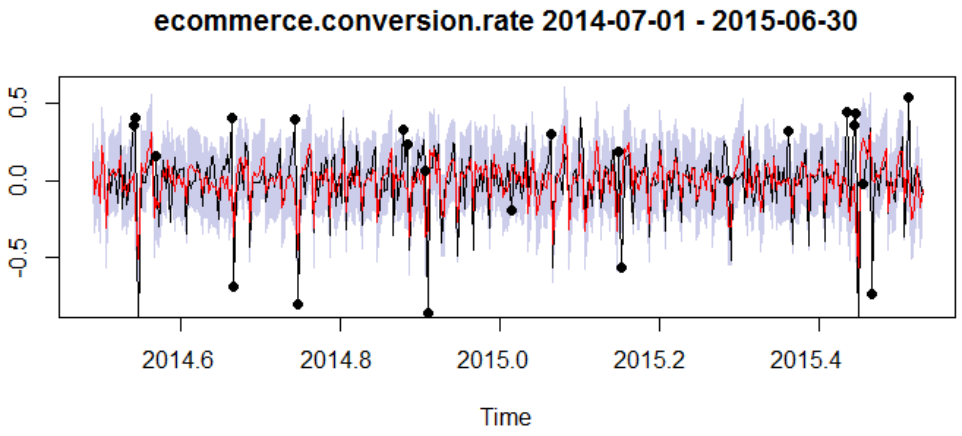


3.26 pav. Kintamojo *pageviews* 2014-07-01 – 2015-06-30 grafiniai rezultatai

Kintamojo *pageviews* pirmųjų metų analizėje buvo rasta 16 anomalijų. Patikrinus jas su tikriausiais duomenimis buvo rastos net 8 tos pačios anomalijos. Lyginant su kitais modeliais, šio kintamojo modelis yra jautrus ir tikslus.

3.17 lentelė. Kintamojo *pageviews* pirmų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	8	5
	False	8	344
Įvertinimai			
Tikslumas (angl. *precision*)		61.5%	
Specifiškumas		61.5%	

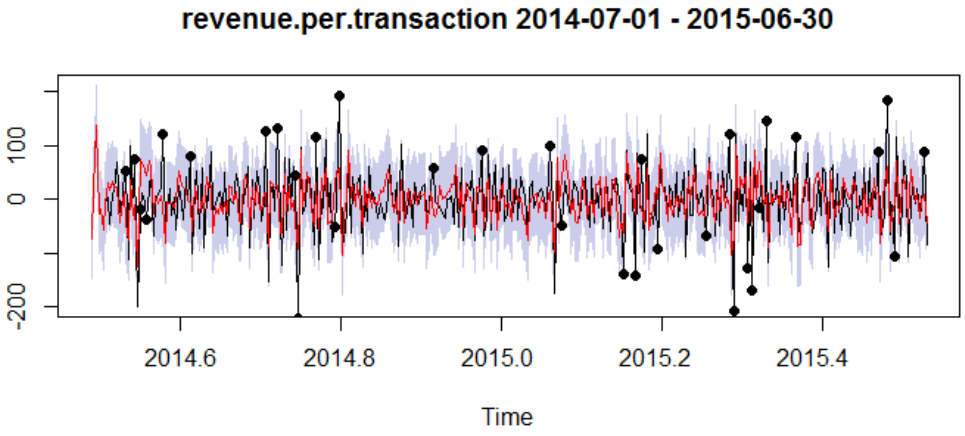


3.27 pav. Kintamojo *ecommerce.conversion.rate* 2014-07-01 – 2015-06-30 grafiniai rezultatai

Ketvirtojo kintamojo VAR(7) modelio analizėje buvo atrastos 25 anomalijos, tačiau patikrinus jas su tikrosiomis reikšmėmis paaiškėjo, kad tik 4 iš jų yra iš tikrųjų anomalijos (30.8 % tikslumas).

3.18 lentelė. Kintamojo *ecommerce.conversion.rate* pirmų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	4	9
	False	21	331
Įvertinimai			
Tikslumas (angl. *precision*)		30.8%	
Specifiškumas		70%	



3.28 pav. Kintamojo *revenue.per.transaction* 2014-07-01 – 2015-06-30 grafiniai rezultatai

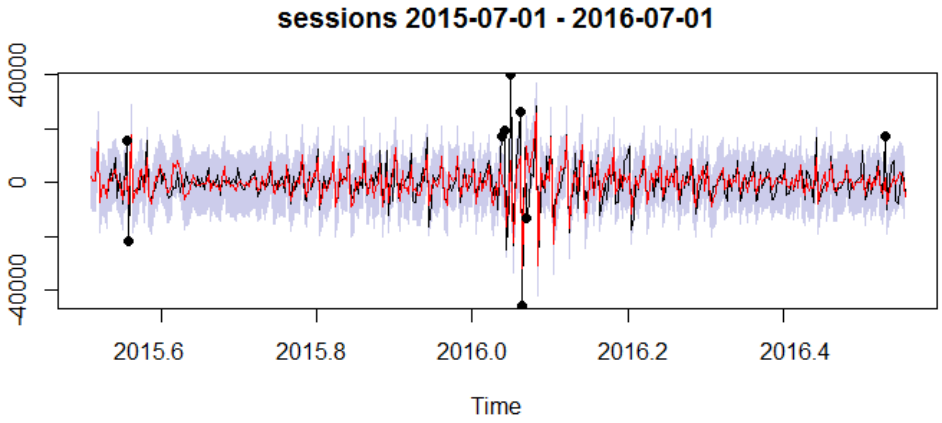
Paskutinio, penktojo, kintamojo analizėje buvo rastas taip pat didelis skaičius anomalijų, net 33. Tačiau patikrinus jas su tikrosiomis reikšmėmis, paaiškėjo, kad tik 1 iš jų sutapo. Tai rodo modelio prastą anomalijų aptikimą ir mažą tikslumą. Kaip ir daugumai, modeliui yra būdingas aukštas specifiškumas.

3.19 lentelė. Kintamojo *revenue.per.transaction* pirmų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	1	12
	False	32	320
Įvertinimai			
Tikslumas (angl. *precision*)		7.7%	
Specifiškumas		72.7%	

3.3.2. Kintamųjų rezultatai antraisiais metais (2015-07-01 – 2016-07-01)

Atlikus automatinę VAR modelio parametro nustatymo funkciją antriesiems metams, buvo gauta, kad kintamiesiems bus naudojamas modelis VAR(8) - bus naudojama 8 dienų lago vektorinė autoregresijos analizė.

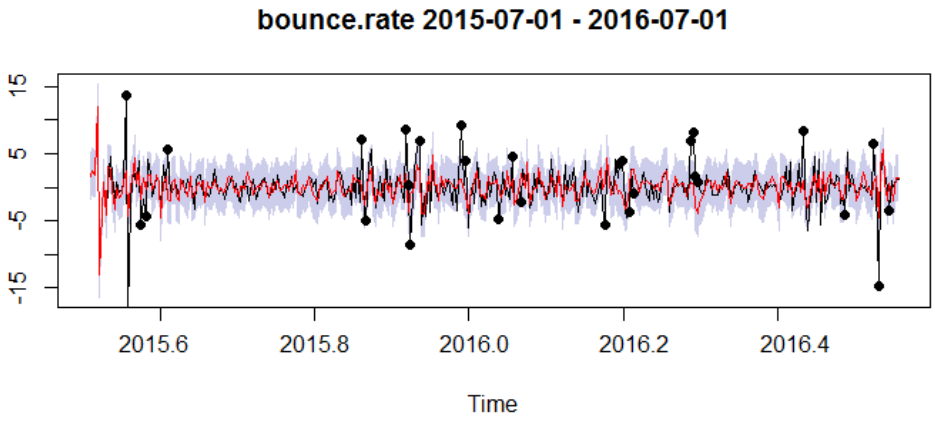


3.29 pav. Kintamojo *sessions* 2015-07-01 – 2016-07-01 grafiniai rezultatai

Atlikta kintamojo *sessions* analizė rado 9 anomalijas per 2015-07-01 – 2016-07-01 laikotarpį. Patikrinus anomalijas su realiomis reikšmėmis, buvo rasta tik 1 anomalija. Sudarytas kintamojo modelis yra labai mažo tikslumo.

3.20 lentelė. Kintamojo *sessions* antrų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	1	12
	False	8	346
Įvertinimai			
Tikslumas (angl. *precision*)		7.7%	
Specifiškumas		40%	

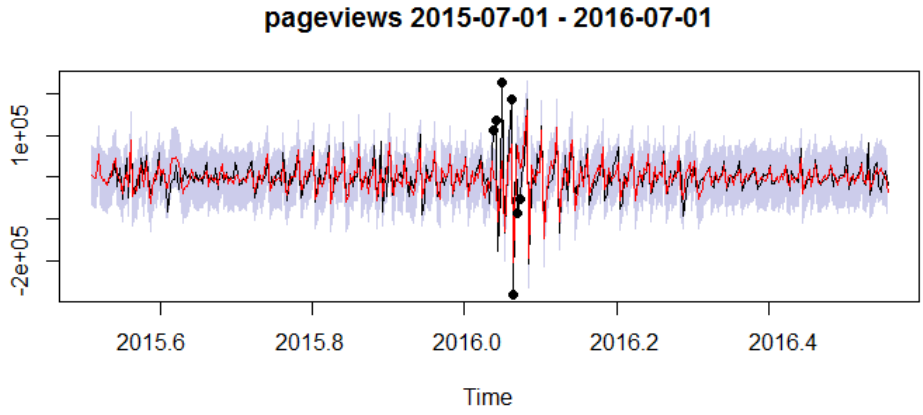


3.30 pav. Kintamojo *bounce.rate* 2015-07-01 – 2016-07-01 grafiniai rezultatai

Kito kintamojo analizėje buvo rastos 30 anomalijų. Patikrinus ar jos yra ir tikruosiuose duomenyse, paaiškėjo, kad tik 5 iš jų buvo tikros. Modelis yra didelio specifiškumo ir mažo tikslumo.

3.21 lentelė. Kintamojo *bounce.rate* antrų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	5	8
	False	25	329
Įvertinimai			
Tikslumas (angl. *precision*)		38.5%	
Specifiškumas		75.8%	

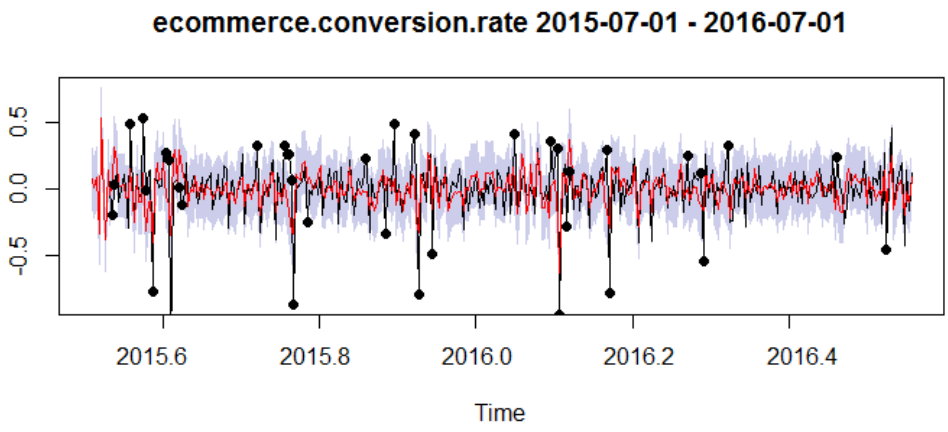


3.31 pav. Kintamojo *pageviews* 2015-07-01 – 2016-07-01 grafiniai rezultatai

Kintamojo *pageviews* antrųjų metų laikotarpiu buvo aptiktos tik 7 anomalijos. Tai parodo, kad per šį laikotarpį buvo mažai svyravimų kintamajame. Nei viena iš rastų anomalijų nebuvo tikruose duomenyse. Tiesa, šiam modeliui nėra būdingos ir pirmo tipo klasifikavimo klaidos, kas yra būdinga daugumai sudarytų modelių (35 %).

3.22 lentelė. Kintamojo *pageviews* antrų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	0	13
	False	7	347
Įvertinimai			
Tikslumas (angl. *precision*)		0%	
Specifiškumas		35%	

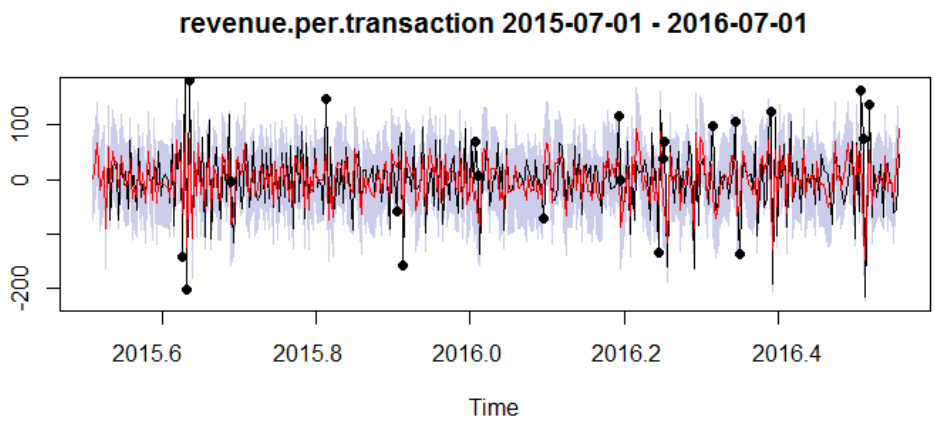


3.32 pav. Kintamojo *ecommerce.conversion.rate* 2015-07-01 – 2016-07-01 grafiniai rezultatai

Kintamojo *ecommerce.conversion.rate* laiko eilutėje antraisiais metais buvo rastos net 38 anomalijos, kas parodo dažnus ir netikėtus svyravimus. Patikrinus anomalijas su tikromis reikšmėmis, buvo nustatyta, kad modelis aptiko tik 2 anomalijas. Tai rodo, kad modelis yra labai netikslus ir labai didelis skaičius klaidų yra pirmo tipo (aukštas specifiškumas).

3.23 lentelė. Kintamojo *ecommerce.conversion.rate* antrų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	2	11
	False	36	318
Įvertinimai			
Tikslumas (angl. *precision*)		15.4%	
Specifiškumas		76.6%	



3.33 pav. Kintamojo *revenue.per.transaction* 2015-07-01 – 2016-07-01 grafiniai rezultatai

Paskutiniojo kintamojo *revenue.per.transaction* laiko eilutėje buvo nustatytos 23 taškai, kurie peržengė pasikliautinumo intervalą. Patikrinus rezultatus su tikromis anomalijomis, buvo nustatyta, kad tik 1 anomalija buvo nustatyta.

3.24 lentelė. Kintamojo *revenue.per.transaction* antrų metų sumaišymo matrica, jautrumo ir tikslumo įvertinimai

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	1	12
	False	22	332
Įvertinimai			
Tikslumas (angl. *precision*)		7.7%	
Specifiškumas		64.7%	

Atlikti modeliavimo su VAR metodika rezultatai buvo vidutiniški. Sudėjus visų kintamųjų rezultatus gauta, kad buvo atpažintos net 176 anomalijos per 2 metų laikotarpį. Tačiau taip pat buvo patikrinta kiek iš šių anomalijų sutampa su tikrosiomis pagal duomenis. Atlikus tai paaiškėjo, kad buvo aptiktos 17 ir 26 anomalijų (65.4 % tikslumas). Daugumai kintamųjų buvo būdingas aukštas specifiškumas, todėl sudėjus rezultatus jis taip pat išliko. Tai rodo, kad kintamieji atpažino skirtingas anomalijas, tačiau tik mažas kiekis jų pasirodė esančios anomalijos ištikrųjų.

Palyginus rezultatus su skirtingais pasikliautinaisiais intervalais (3.26. ir 3.27. lent.) parodė, kad intervalo plėtimas iki 99% neaptiko tik 2 anomalijomis mažiau negu 95%. Tačiau daugiau negu

dvigubai sumažėjo neteisingai klasifikuotų anomalijų skaičius. Sumažinus pasikliovimo intervalo dydį iki 90%, buvo rasta tik 3 anomalijomis daugiau, tačiau neteisingai identifikuotų anomalijų skaičius išaugo iki 240 taškų.

3.25 lentelė. Visų kintamųjų bendra sumaišymo matrica, jautrumo ir tikslumo įvertinimai su 95% pasikliovimo intervalu

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	17	9
	False	159	547
Įvertinimai			
Tikslumas (angl. *precision*)		65.4%	
Specifiškumas		94.6%	

3.26 lentelė. Visų kintamųjų bendra sumaišymo matrica, jautrumo ir tikslumo įvertinimai su 90% pasikliovimo intervalu

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	20	6
	False	240	466
Įvertinimai			
Tikslumas (angl. *precision*)		76.9%	
Specifiškumas		97.6%	

3.27 lentelė. Visų kintamųjų bendra sumaišymo matrica, jautrumo ir tikslumo įvertinimai su 99% pasikliovimo intervalu

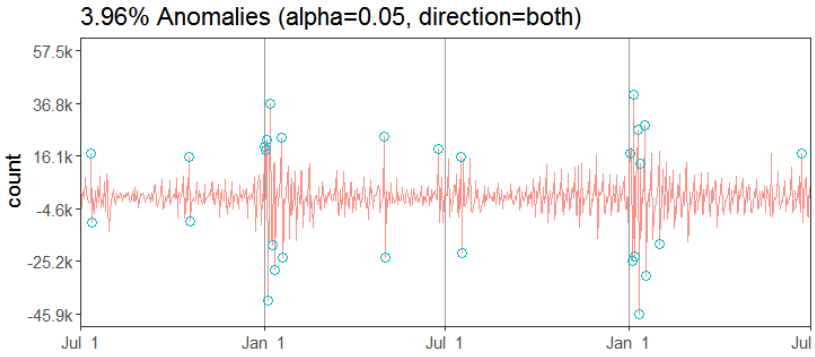
		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	15	11
	False	76	630
Įvertinimai			
Tikslumas (angl. *precision*)		57.7%	
Specifiškumas		87.4%	

3.4. Anomalijų aptikimo rezultatai naudojant „AnomalyDetection“ paketą

„AnomalyDetection“ yra vienmačių laiko eilučių anomalijų aptikimo paketas. Todėl bus sudaryti 5 modeliai su kiekvienu kintamuoju. Vėliau bus sudėti rezultatai ir patikrinta kiek anomalijų

pavyko aptikti kartu panaudojant visus kintamuosius. Visiems kintamiesiems buvo naudoti tie patys parametrai: maksimalus anomalijų skaičius (*max_anoms*) - 0.04 (nes iš viso duomenyse yra 3,5% anomalijų), anomalijų kryptis (*direction*) parinkta dvipusė (*both*), kad būtų rasti ir staigūs šuoliai ir kritimai, įjungta grafiniai rezultatai (*plot=TRUE*), kad būtų galima matyti anomalijas. Kaip ir ankstesniuose modeliuose, nors tyrime buvo naudoti 3 pasikliovimo intervali (90%, 95%, 99%), tačiau atskirtų kintamųjų atveju bus pateikti tik 95% intervalo rezultatai. Pabaigoje pateikti visų 3 intervalų rezultatai.

Iš pradžių buvo atliekama sesijų skaičiaus (*sessions*) anomalijų paieška (3.34. pav.).



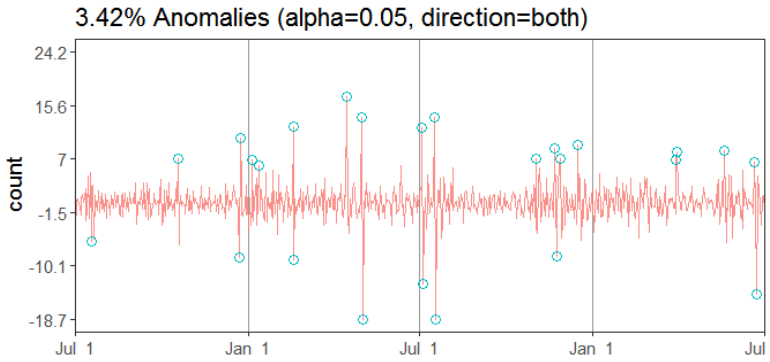
3.34 pav. Kintamojo *sessions* anomalijų aptikimas naudojantis „AnomalyDetection“ paketą

Atlikus anomalijų paiešką, buvo rastos 28 anomalijos(pažymėtos apskritimais paveiksle). Tačiau patikrinus su realiais duomenimis sumaišymų matricos pagalba buvo atpažintos 8 anomalijos (3.28. lent.).

3.28 lentelė. Kintamojo *sessions* sumaišymo matrica, jautrumo ir tikslumo įvertinimai naudojant “AnomalyDetection” paketą

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	8	18
	False	20	688
Įvertinimai			
Tikslumas (angl. *precision*)		30.8%	
Specifiškumas		52.6%	

Antrojo kintamojo *bounce.rate* anomalijų paieškos rezultatai grafiniu būdu (3.35. pav.).



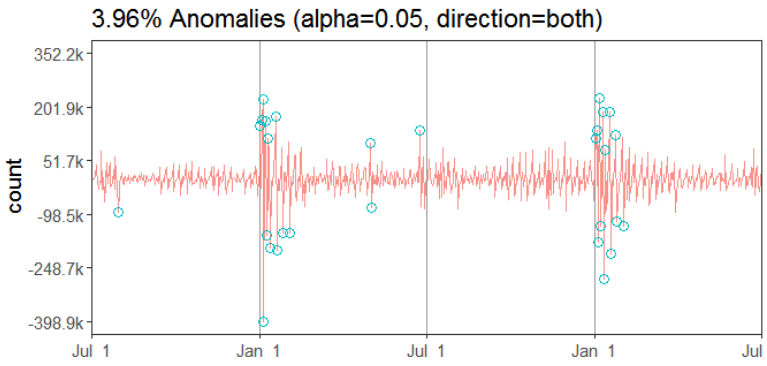
3.35 pav. Kintamojo *bounce.rate* anomalijų aptikimas naudojantis „AnomalyDetection“ paketą

Atlikus anomalijų paiešką, buvo rastos 25 anomalijos. Tačiau patikrinus su realiais duomenimis sumaišymų matricos pagalba buvo atpažintos 6 anomalijos. (3.29. lent.). Modelis yra mažo tikslumo, modelio klaidų tipai yra pasiskirstę panašiai, todėl išskirti vieno klaidų tipo negalime.

3.29 lentelė. Kintamojo *bounce.rate* sumaišymo matrica, jautrumo ir tikslumo įvertinimai naudojant “AnomalyDetection” paketą

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	6	20
	False	19	687
Įvertinimai			
Tikslumas (angl. *precision*)		23.1%	
Specifiškumas		48.7%	

Kintamojo *pageviews* anomalijų paieškos rezultatai grafiniu būdu (3.36. pav.).



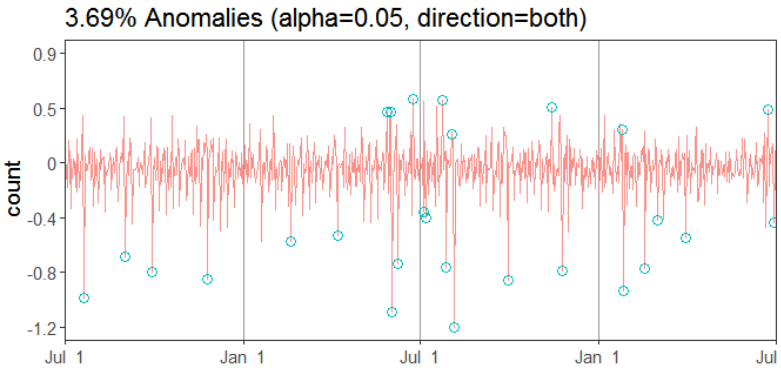
3.36 pav. Kintamojo *pageviews* anomalijų aptikimas naudojantis „AnomalyDetection“ paketą

Atlikus anomalijų paiešką, buvo rastos 29 anomalijos. Tačiau patikrinus su realiais duomenimis sumaišymų matricos pagalba buvo atpažintos 7 anomalijos (3.30. lent.). Vertinant įverčius, galima sakyti, kad modeliui trūksta tikslumo. Specifiškumo rodiklis parodo, kad sunku įvardinti kodėl modelis nėra tikslus, nes abiejų tipų klasifikavimo klaidos yra panašaus skaičiaus.

3.30 lentelė. Kintamojo *pageviews* sumaišymo matrica, jautrumo ir tikslumo įvertinimai naudojant “AnomalyDetection” paketą

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	7	19
	False	22	684
Įvertinimai			
Tikslumas (angl. *precision*)		26.9%	
Specifiškumas		53.7%	

Kintamojo *ecommerce.conversion.rate* anomalijų aptikimo rezultatai grafiniu būdu (3.37. pav.).



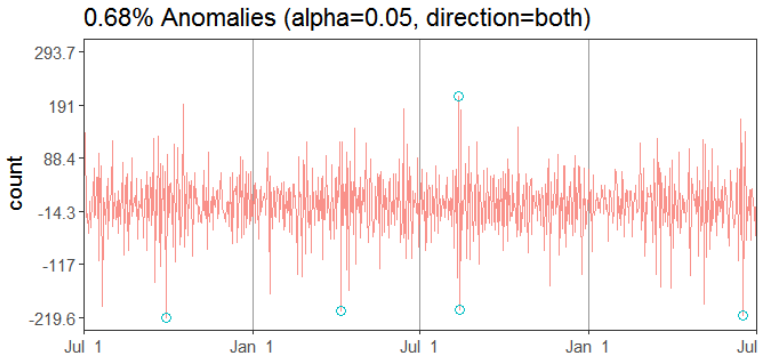
3.37 pav. Kintamojo *ecommerce.conversion.rate* anomalijų aptikimas naudojantis „AnomalyDetection“ paketą

Atlikus anomalijų paiešką, buvo rastos 27 anomalijos. Tačiau patikrinus su realiais duomenimis sumaišymų matricos pagalba buvo atpažintos tik 4 anomalijos (3.31. lent.). Šio kintamojo modelis yra laba mažo tikslumo (15.4 %).

3.31 lentelė. Kintamojo *ecommerce.conversion.rate* sumaišymo matrica, jautrumo ir tikslumo įvertinimai naudojant “AnomalyDetection” paketą

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	4	22
	False	23	683
Įvertinimai			
Tikslumas (angl. *precision*)		15.4%	
Specifiškumas		51.1%	

Paskutinio kintamojo, *revenue.per.transaction*, anomalijų atpažinimo rezultatai grafiniu būdu. (3.38. pav.).



3.38 pav. Kintamojo *revenue.per.transaction* anomalijų aptikimas naudojantis „AnomalyDetection“ paketą

Atlikus anomalijų paiešką su paskutiniu kintamuoju, buvo rasta mažiausiai anomalijų - 5. Tačiau patikrinus su realiais duomenimis sumaišymų matricos pagalba buvo atmestos visos rastos išskirtys (3.32. lent.). Šiam modeliui yra būdingos antro tipo klasifikavimo klaidos – daugumą kintamųjų priskiria normaliems duomenims.

3.32 lentelė. Kintamojo *revenue.per.transaction* sumaišymo matrica, jautrumo ir tikslumo įvertinimai naudojant "AnomalyDetection" paketą

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	0	26
	False	5	701
Įvertinimai			
Tikslumas (angl. *precision*)		0%	
Specifiškumas		16.1%	

Sudėjus visų kintamųjų atpažintas anomalijas ir pašalinus pasikartojančias eilutes buvo atpažintos 78 anomalijos. Tačiau tik 14 iš jų sutapo su tikrųjų duomenų anomalijomis.(3.33. lent.). Tai parodo tik 17,9 % tikslumą. Modeliui būdingos pirmo tipo klasifikavimo klaidos (dideli specifiškumas).

3.33 lentelė. Visų kintamųjų sumaišymo matrica, jautrumo ir tikslumo įvertinimai naudojant "AnomalyDetection" paketą ir 95 % pasikliovimo intervalą

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	14	12
	False	64	642
Įvertinimai			
Tikslumas (angl. *precision*)		53.8%	
Specifiškumas		84.2%	

3.34 lentelė. Visų kintamųjų sumaišymo matrica, jautrumo ir tikslumo įvertinimai naudojant "AnomalyDetection" paketą ir 90 % pasikliovimo intervalą

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	14	12
	False	68	638
Įvertinimai			
Tikslumas (angl. *precision*)		53.8%	
Specifiškumas		85.0%	

3.35 lentelė. Visų kintamųjų sumaišymo matrica, jautrumo ir tikslumo įvertinimai naudojant "AnomalyDetection" paketą ir 99 % pasikliovimo intervalą

		Prognozuojamos reikšmės	
		True	**False**
Tikrosios reikšmės	**True**	13	13
	False	59	466
Įvertinimai			
Tikslumas (angl. *precision*)		50.0%	
Specifiškumas		81.9%	

3.5. Modelių rezultatų palyginimas

Bandant aptikti anomalijas iš duomenų buvo naudojami trys metodai – autoregresijos su slenkančiu vidurkiu ir išoriniais kintamaisiais (ARIMA), vektorinės autoregresijos modelis su išoriniais kintamaisiais (VAR) modeliai ir „AnomalyDetection" R paketu.

Modelius galima lyginti pagal kelis kriterijus – atrastų anomalijų skaičių, tikslumą, santykį tarp jų rastų anomalijų ir tikrųjų anomalijų, palyginus rezultatus su tikraisiais duomenimis.

Tyrimo metu buvo išsiaiškinta, kad daugiausiai anomalijų atpažinta remiantis VAR laiko eilutės modeliavimo metodu. Jo pagalba buvo aptiktos 15 (99% paskliovimo intervalas) 17 (95 % pasikliovimo intervalas) ir 20 (90 % pasikliovimo intervalas) iš 26 anomalijų. Mažiausiai anomalijų aptiko "AnomalyDetection" R paketas. Jis atpažino vienodą skaičių anomalijų – 14 (53.8 % tikslumas) su 95 % ir 90 % pasikliovimo intervalu. Tiesa, su 99 % pasiliovimo intervalu buvo atpažinta 13 anomalijų. Visiems modeliams yra būdingas labai aukštas specifiškumas. Tai rodo, kad modeliai randa daug išskirčių, tačiau tik maža dalis iš jų yra anomalijos. Vieno sprendimo, kaip padidinti ir tikslumą ir sumažinti specifiškumo tikriausiai nėra. Tačiau norint tai įvykdyti reikia turėti

daugiau kintamųjų, kurie suteiktų papildomos informacijos. Pasirinkti pasiliovimo intervalą tyrime reikėtų priklausomai nuo tyrimo tikslo. Jei yra norima atrasti kuo daugiau anomalijų – patartina rinktis 99 %, tačiau jei yra akcentuojamas tikslumas ir norima kuo tiksliau rasti anomalijas, geriau rinktis mažesnį – 90% intervalą.

IŠVADOS

1. Atlikus literatūros analizę apie anomalijas ir elektroninę prekybą galima įsitikinti, kad anomalijų aptikimas padėtų bendrovėms geriau pažinti savo pirkėjus.
2. Atlikus duomenų analizę buvo parinkti autoregresijos su slenkančiu vidurkiu ir išoriniais veiksniais (ARIMA), vektorinės autoregresijos su išoriniais veiksniais (VAR) ir R paketu „AnomalyDetection“.
3. Duomenų rinkinį sudarė 9 kintamieji: *ga.date, date, sessions, bounce,rate,pageviews,ecommerce.convertion.rate,revenue.per.transaction,significant, reason*. Duomenų rinkinyje yra 733 eilutės, tarp kurių yra 26 anomalijos. Dėl pasikartojančios ar jokios naujos informacijos neturėjimo buvo pašalinti kintamieji *ga.date*(kartojasi su *date*) ir *reason*. Atlikus ADF, Box – Ljung testus ir iš grafinių autokoreliacijos (ACF) ir dalinės autokoreliacijos (PACF) funkcijų buvo nustatyta, kad duomenys yra nestacionarūs, todėl jie buvo diferencijuojami. Po diferencijavimo paaiškėjo, kad duomenims taip pat yra būdingas savaitinis sezoniškumas. Prieš sudarant modelius buvo sukurtas naujas kintamasis *day*, kuris nurodė savaitės dieną.
4. Autoregresijos su slenkančiu vidurkiu ir išoriniais veiksniais (ARIMA) metodu buvo aptikta 16 (95 % pasikliovimo intervalas), 18 (90 % pasikliovimo intervalas) ir 14 (99 % pasikliovimo intervalas) anomalijų iš 26, o tai reiškia, kad praktikoje šis modelis vidutiniškai aptinka anomalijas iš el. prekybos statistikos duomenų.
5. Vektorinės autoregresijos su išoriniais veiksniais (VAR) metodu buvo rasta 17 (95 % pasikliovimo intervalas), 20 (90 % pasikliovimo intervalas) ir 15 (99 % pasikliovimo intervalas) anomalijų iš 26, o tai reiškia praktikoje, kad šis modelis vidutiniškai aptinka anomalijas pagal el. prekybos statistikos duomenų.
6. R paketu „AnomalyDetection“ paketu metodu su pasikliovimo intervalais 95 ir 90 % buvo atpažinta 14 anomalijų (53.8 %), o su 99 % - 13 Tai reiškia praktikoje paketas nėra labai veiksmingas aptikti anomalijas daugiadimensinėse laiko eilutėse.
7. Remiantis gautais rezultatais, buvo nuspręsta, kad geriausiai anomalijas aptinka vektorinės autoregresijos su išoriniais veiksniais metodas, tačiau vertinant visus modelius, anomalijų aptikimo rezultatai yra vidutiniški. Visiems modeliams yra būdingas aukštas pirmo tipo klasifikavimo klaidos. Pasikliovimo intervalą tiriant duomenis reikėtų

pasirinkti remiantis tuo ar norima atrasti kuo daugiau anomalijų ar yra svarbu klaidingai identifikuotų taškų skaičius.

LITERATŪROS SĄRAŠAS

1. NIRANJANAMURTHY, M., ir kt. Analysis of e-commerce and m-commerce: advantages, limitations and security issues.[interaktyvus*] International Journal of Advanced Research in Computer and Communication Engineering*, 2013, 2.6. [žiūrėta: 2017-04-28]. Prieiga per internetą: http://www.academia.edu/download/33193840/7-Niranjanamurthy-Analysis_of_E-Commerce_and_M-Commerce_Advantages.pdf
2. BAKANAUSKAS, ARVYDAS; LIESIONIS, VYTAUTAS. Elektroninis marketingas. VDU leidykla, 2008.
3. LIU, Jianjia; LU, Lei. Development of E-commerce Statistics and the Implications. In: Proceedings of the 7th international conference on Electronic commerce. ACM, 2005. p. 70-72. [žiūrėta 2017-05-01]. Prieiga per internetą: http://dl.acm.org/ft_gateway.cfm?id=1089567&ftid=328940&dwn=1&CFID=741622205&CFTOKEN=44206871
4. BARREIRA, João, et al. Analysis, Specification and Design of an e-Commerce Platform That Supports Live Product Customization.[interaktyvus*] In: International Conference on Software Process Improvement. Springer International Publishing*, 2016. p. 267-274. [žiūrėta 2017-04-18] Prieiga per internetą: https://www.researchgate.net/profile/Jose_Martins23/publication/308960197_Analysis_Specification_and_Design_of_an_e-Commerce_Platform_That_Supports_Live_Product_Customization/links/5804d7e608aef87fbf3ba258.pdf
5. Lietuvos statistikos departamentas (2013). Asmenys, pirkę ar užsakę prekių ar paslaugų internetu [žiūrėta 2017-04-20]. Prieiga per internetą http://osp-old.stat.gov.lt/web/guest/statistiniu-rodikliu-analize?portletFormName=visualization&hash=6f43f70b-1503-48b2-a402-30c46fff33a6.
6. Eurostat(2017). Individuals using the internet for ordering goods or services [interaktyvus][žiūrėta 2017-03-15]. Prieiga per internetą: http://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tin00096&plugin=1
7. BASARIR-OZEL, B; MARDIKYA, S. Factors affecting E-commerce adoption: A case of Turkey. International Journal of Management Science & Technology Information. 23, 1-11, Jan. 2017. ISSN: 19230265.[žiūrėta 2017-04-15]. Prieiga per internetą: http://web.b.ebscohost.com/ehost/detail/detail?vid=7&sid=d24ce7d1-7b27-43b2-8ffa-3923f7c2ed32%40sessionmgr103&hid=124&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#
8. CHEN, Hsinchun; CHIANG, Roger HL; STOREY, Veda C. Business intelligence and analytics: From big data to big impact [interaktyvus]. *MIS quarterly*, 2012, 36.4: 1165-1188.[žiūrėta 2017-03-25]. Prieiga

per internetą: http://www.academia.edu/download/32970305/FROM_BIG_DATA_TO_BIG_IMPACT.pdf

9. DAVIDAVIČIENĖ, Vida, ir kt. Elektroninės prekybos interneto svetainių Lietuvoje vertinimas[interaktyvus]. *Informacijos mokslai*, 2011, 55: 103-116. [žiūrėta 2017-03-02].Prieiga per internetą: http://www.journals.vu.lt/informacijos-mokslai/article/download/3164/2282
10. POWELL, Guy R. *Marketing calculator: Measuring and managing return on marketing investment*. John Wiley & Sons, 2012.
11. DEMERS, Jayson. 10 Online Marketing Metrics You Need to Be Measuring. *Forbes, August*, 2014, 15.[žiūrėta 2017-04-25]. Prieiga per internetą: https://www.forbes.com/sites/jaysondemers/2014/08/15/10-online-marketing-metrics-you-need-to-be-measuring/#467115ee76c1
12. HASAN, Layla; MORRIS, Anne; PROBETS, Steve. Using Google Analytics to evaluate the usability of e-commerce sites. Human centered design, 2009, 697-706.[žiūrėta: 2017-03-23]. Prieiga per internetą: https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/5685/1/Using%20Google
13. CHANDOLA, Varun; BANERJEE, Arindam; KUMAR, Vipin. Anomaly detection: A survey [interaktyvus]. *ACM computing surveys (CSUR)*, 2009, 41.3: 15 [žiūrėta 2017-04-18]. Priega per internetą: https://pdfs.semanticscholar.org/7b5a/c1fb5627addf92ad5804a6569a6cfa9385ac.pdf
14. HODGE, Victoria J.; LEES, Ken J.; AUSTIN, James L. A high performance k-NN approach using binary neural networks. [interaktyvus] *Neural Networks*, 2004, 17.3: 441-458. [žiūrėta 2017-03-10]. Prieiga per internetą: http://eprints.whiterose.ac.uk/768/1/hodgevj5.pdf
15. SMITH, Michael R.; MARTINEZ, Tony. Improving classification accuracy by identifying and removing instances that should be misclassified.[interaktyvus] In: *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, 2011. p. 2690-2697. [žiūrėta 2017-03-14]. Prieiga per internetą: http://www.academia.edu/download/31723952/smith.ijcnn2011.pdf
16. HYNDMAN, Rob J.; WANG, Earo; LAPTEV, Nikolay. Large-scale unusual time series detection. In: Data Mining Workshop (ICDMW),[interaktyvus] 2015 IEEE International Conference on. IEEE, 2015. p. 1616-1619 [žiūrėta 20174-04-05]. Prieiga per internetą: http://robjhyndman.com/papers/icdm2015.pdf
17. Imdadullah, Muhammad.. "Time Series Analysis". Basic Statistics and Data Analysis. itfeature.com. Retrieved 2 January 2014. [žiūrėta 2017-04-01]. Prieiga per internetą: http://itfeature.com/time-series-analysis-and-forecasting/time-series-analysis-forecasting
18. CARRION-I-SILVESTRE, Josep Lluís; SANSÓ, Andreu. A guide to the computation of stationarity tests. Empirical Economics, 2006, 31.2: 433. [žiūrėta 2017-04-03]. Prieiga per internetą: http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=03777332&AN=21467255&h=sohHc9YNTCaHo5T9q7zLqtny1UlDB8DBW5POLmman7hcGImXNTVy8%2F9ABfaiWvXv2oAfmyLknpEHepejAzODDQ%3D%3D&crl=f

19. BOGUSLAUSKAS, V. Ekonometrija: mokomoji knyga. Kaunas: Technologija, 2010.

20. SUN, Wei. Business Cycle Synchronization and Monetary Policy Coordination Between the US and China: Evidence From a Structural VAR Model. The Chinese Economy, 2017, 50.1: 3-20. [žiūrėta 2017-03-20]. Prieiga per internetą: SUN, Wei. Business Cycle Synchronization and Monetary Policy Coordination Between the US and China: Evidence From a Structural VAR Model. *The Chinese Economy*, 2017, 50.1: 3-20.

21. KARTHIKA, M., et al. Forecasting of meteorological drought using ARIMA model. Indian Journal of Agricultural Research, 2017, 51.2. [žiūrėta 2017-03-22] Prieiga per internetą: http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=03678245&AN=122739502&h=4x0dK0gaNkvobgFlSKMaul%2FBVSyX%2FZKkZyVACqcjByE4h2v7mItfQW56nSAJh9cpktDAOdIj%2BQMlR3QIhE5c2A%3D%3D&crl=f

22. SHUMWAY, Robert H.; STOFFER, David S. *Time series analysis and its applications: with R examples*[interaktyvus]. Springer Science & Business Media, 2010 [žiūrėta 2017-04-10]. Prieiga per internetą: http://www.stat.pitt.edu/stoffer/tsa4/tsaEZ.pdf

23. CHO, Vincent. A comparison of three different approaches to tourist arrival forecasting. Tourism management, 2003, 24.3: 323-330. [žiūrėta 2017-04-10]. Prieiga per internetą: https://scholar.google.lt/scholar?output=instlink&q=info:zaLv6fuEdz8J:scholar.google.com/&hl=lt&as_sdt=0,5&scillfp=17755385224248631 09&oi=lle

24. Anomaly Detection R pachage.[interaktyvus] 2015 [žiūrėta 2017-04-01]. Prieiga per internetą: https://github.com/twitter/AnomalyDetection