

ნაუქასტინგი, მაღალი სიხშირის მონაცემების, გუგლ ტრენდებისა და მანქანური სწავლების გამოყენებით

მიხეილ მღებრიშვილი

კვლევით ნაშრომი

WP 01/2022

კვლევითი ნაშრომების სერია წარმოადგენს ავტორ(ებ)ის მიმდინარე კვლევას და მიზნად ისახავს კომენტარების მიღებას და მსჯელობის წახალისებას. ნაშრომში გამოთქმული მოსაზრებები ასახავს ავტორ(ებ)ის აზრს და არ წარმოადგენს საქართველოს ფინანსთა სამინისტროს ოფიციალურ პოზიციას.

ნაუქასთინგი, მაღალი სიხშირის მონაცემების, გუგლ ტრენდებისა და მანქანური სწავლების გამოყენებით¹

მიხეილ მღებრიშვილი²

დეკემბერი, 2022

კვლევითი ნაშრომების სერია წარმოადგენს ავტორ(ებ)ის მიმდინარე კვლევას და მიზნად ისახავს კომენტარების მიღებას და მსჯელობის წახალისებას. ნაშრომში გამოთქმული მოსაზრებები ასახავს ავტორ(ებ)ის აზრს და არ წარმოადგენს საქართველოს ფინანსთა სამინისტროს ოფიციალურ პოზიციას.

აბსტრაქტი

მიმდინარე ეკონომიკურ ტენდენციებზე დაკვირვება წარმოადგენს სწორი და დროული ეკონომიკური პოლიტიკის გატარების ერთ-ერთ შემადგენელ ნაწილს. ძირითადად, ეკონომიკური ინდიკატორები, მათ შორის მთლიანი შიდა პროდუქტის (მშპ-ს) მონაცემები რამდენიმეკვირიანი ან რამდენიმეთვიანი დაგვიანებით ქვეყნდება სტატისტიკის ეროვნული სამსახურის მიერ. ამ საინფორმაციო ვაკუუმის შესავსებად ნაშრომი გვთავაზობს ნაუქასთინგის ტრადიციული და თანამედროვე მეთოდების სწრაფი სიხშირის მონაცემებს, მათ შორის Google Trend-ის საძიებო კატეგორიების ინდექსებს. კვლევის დროს მიღებული შედეგების მიხედვით, მანქანური სწავლების მეთოდები შედარებით მაღალი სიზუსტით გამოირჩევა, ვიდრე ტრადიციული მეთოდები. მანქანური სწავლების ალგორითმები, როგორებიცაა Lasso, Ridge, SVR, Linear SVM, Neural Network საშუალოდ 17-21 პროცენტით უკეთეს შედეგს იძლევა, ვიდრე ტრადიციული დინამიკური ფაქტორული მოდელი. როგორც დინამიკური ფაქტორული მოდელი, ასევე მანქანური სწავლების მეთოდები, უმეტეს შემთხვევაში სწორად აფასებს ეკონომიკური აქტივობის ტრეკტორიას. ასევე, აღნიშნული მოდელები კარგად აფიქსირებენ მობრუნების წერტილებს, განსაკუთრებით კი ისეთ შემთხვევებში, როდესაც მყისიერი ვარდნა ან ზრდა ხდება, რაც დამახასიათებელი იყო COVID-19-ის პერიოდისათვის.

JEL codes: C13, C32, C38, C45, C51, C52, C53, C61

საკვანძო სიტყვები: ნაუქასთინგი, გუგლის მონაცემები, გუგლ ტრენდები, მაღალი სიხშირის მონაცემები, მანქანური სწავლება.

¹ ავტორი მაღლობას უხდის სამინისტროს მაკროეკონომიკური ანალიზისა და ფისკალური პოლიტიკის დაგეგმვის დეპარტამენტის თანამშრომლებს რეკომენდაციებისა და უკუკავშირისთვის.

² საქართველოს ფინანსთა სამინისტროს მაკროეკონომიკური ანალიზისა და ფისკალური პოლიტიკის დაგეგმვის დეპარტამენტის მაკროეკონომიკური მოდელირების განყოფილების მესამე კატეგორიის უფროსი სპეციალისტი (e-mail: m.mgebrishvili@mof.ge)

სარჩევი

არა-ტექნიკური შეჯამება	3
1. შესავალი.....	5
2. ლიტერატურის მიმოხილვა.....	6
3. მონაცემები	11
3.1 მონაცემების თავისებურება საქართველოსთვის	11
3.2 მონაცემების აღწერა და დამუშავება	12
4. მოდელები და მეთოდოლოგია	16
4.1 DFM მეთოდოლოგია	16
4.2 მანქანური სწავლების მეთოდები და მოდელები	17
4.3 მოდელების შერჩევა და ნაუქასთინგის სიზუსტის შეფასება	19
5. შედეგები	22
6. დასკვნა.....	25
7. გამოყენებული ლიტერატურა.....	26
8. დანართი	31
8.1 ნაშრომში გამოყენებული მონაცემები	31
8.2 მანქანური სწავლების მოდელები.....	32
8.2.a რეგულარიზებული წრფივი რეგრესიები (LASSO, Ridge, Elastic Net).....	32
8.2.b კერნელის მეთოდები (Support Vector Machine (SVM)).....	34
8.2.c Ensemble methods (Decision trees, Random Forests, Gradient boosting, Adaptive boosting).....	35
8.2.d ნეირონული ქსელი (Neural Network).....	37
8.3 მოდელების შედეგები.....	39

გრაფიკები

გრაფიკი 1. კვარტალური და ყოველთვიური წლიური მშპ-ს ზრდის ტემპი	11
გრაფიკი 2. კვარტალური და ყოველთვიური მშპ-ს ზრდის ტემპების ურთიერთკავშირი	12
გრაფიკი 3. დროითი მწკრივებისთვის კროს-ვალიდაციის მეთოდები.....	20
გრაფიკი 4. მოდელირების პროცესი	21
გრაფიკი 5. მობრუნების წერტილები.....	24
გრაფიკი 6. მოდელებში გამოყენებული მონაცემები.....	32
გრაფიკი 7. მოდელის შედეგები მეორე ტიპის მონაცემებისთვის.....	39
გრაფიკი 8. მოდელის შედეგები პირველი ტიპის მონაცემებისთვის	40

ცხრილები

ცხრილი 1. მანქანური სწავლების მეთოდების მოკლე აღწერა	17
ცხრილი 2. მიღებული მოდელების შედეგები.	22
ცხრილი 3. ნაშრომში გამოყენებული მონაცემები	31
ცხრილი 4. მეორე ტიპის მონაცემების ნაუქასთინგის სიზუსტის შეფასება ავტორის გაანგარიშებები.....	39
ცხრილი 5. პირველი ტიპის მონაცემების ნაუქასთინგის სიზუსტის შეფასება	40

არა-ტექნიკური შეჯამება

მოცემულ ნაშრომში განხილულია ნაუქასთინგის (რაც გულისხმობს აწმყოს პროგნოზირებას/შეფასებას) სხვადასხვა მეთოდი და მათი პრაქტიკაში გამოყენების შესაძლებლობები. ლიტერატურაში ნაუქასთინგისთვის ყველაზე ხშირად განხილული მეთოდია, ისეთი ტიპის მოდელები, რომლებშიც გამოყენებულია ბევრი ე.წ. სწრაფი ინდიკატორი. აღნიშნული მრავალი ინდიკატორის ერთობლიობიდან გამოიყოფა საერთო მახასიათებლების მქონე რამდენიმე ფაქტორი, რომელიც მთლიანი ერთობლიობის უმეტესი ნაწილის ახსნის საშუალებას იძლევა. ამის შემდგომ, გამოყოფილი/შეფასებული ფაქტორების საშუალებით ხდება მიმდინარე ეკონომიკური აქტივობის (მთლიანი შიდა პროდუქტის ზრდის) შეფასება. ამ მიდგომასთან ერთად სულ უფრო პოპულარული ხდება მანქანური სწავლების მეთოდების გამოყენება ნაუქასთინგისთვის, ვინაიდან, ეს უკანასკნელნი უმეტესწილად პრაქტიკაში სწორედ პროგნოზირების მიზნებისთვის გამოიყენება.

ნაშრომში ტრადიციულ და თანამედროვე მეთოდებთან ერთად გამოყენებულია მონაცემების მოპოვების თანამედროვე წყაროებიც, რომელსაც წარმოადგენს Google Trend-ის ძიების ინდექსები. გუგლის მონაცემების გამოყენება საშუალებას იძლევა მოხდეს დამატებითი მაღალი სიხშირისა და ხარისხის მონაცემების დამატება მოდელირების პროცესში.

საქართველოს მონაცემებით მიღებული შედეგების მიხედვით, თანამედროვე მანქანური სწავლების მეთოდებით ტრადიციულ მეთოდებთან შედარებით უკეთესი შედეგების მიღებაა შესაძლებელი. ამასთან, ნაშრომში შეფასებული მოდელები მაღალი სიზუსტით ახდენენ ეკონომიკური მდგომარეობის მიმართულების შეფასებას, საშუალოდ 70-77 პროცენტ შემთხვევაში მოდელით შეფასებული მიმართულება სწორად ხდება იდენტიფიცირებული. მოდელები ასევე კარგად ართმევენ თავს მობრუნების წერტილების დაფიქსირებას, განსაკუთრებით კი ისეთ პერიოდებში როდესაც ეკონომიკური მდგომარეობა მკვეთრ ცვლილებებს ექვემდებარება, რაც COVID-19 პერიოდისთვის იყო დამახასიათებელი. სწორედ ამ პერიოდისთვის, მოცემულმა მოდელებმა კარგად გაართვეს თავი მყისიერი ეკონომიკური ვარდნის დაფიქსირებას, ისევე როგორც 2021 წლის მარტში ეკონომიკური ზრდის ნახტომისებურ გაუმჯობესებას.

საშუალოდ, მანქანური სწავლების მეთოდები 17-21 პროცენტით უკეთეს შედეგს იძლევიან ვიდრე ეს ტრადიციული მეთოდით მიიღწევა. მიღებული შედეგები თანხვედრაშია გავცელებულ ლიტერატურასთან და სხვა ქვეყნის მონაცემებზე დაფუძნებულ კვლევებთან.

1. შესავალი

მიმდინარე ეკონომიკურ ტენდენციებზე დაკვირვება წარმოადგენს სწორი და დროული ეკონომიკური პოლიტიკის გატარების ერთ-ერთ შემადგენელ ნაწილს, განსაკუთრებით კი ისეთ კრიზისულ პერიოდებში როგორც COVID-19-ით გამოწვეული კრიზისი იყო, რის დროსაც საჭირო გახდა მყისიერი გადაწყვეტილებების მიღება. მსოფლიოს მრავალ ქვეყანაში ძირითადი ეკონომიკური ინდიკატორები, მათ შორის მშპ-ს მონაცემები რამდენიმე კვირიანი ან თვიანი დაგვიანებით ქვეყნდება სტატისტიკის სამსახურების მიერ, ამიტომ იმ დროს როცა ეკონომიკური სიტუაცია მკვეთრად იცვლება ამას პოლიტიკის გამტარებლები დაგვიანებით პასუხობენ. ეკონომიკური ინდიკატორების პროგნოზირება მიმდინარე პერიოდში (შემდეგში ნაუქასთინგი, nowcasting), სწორედ ამ პრობლემის გადაჭრაში გვეხმარება. ის საშუალებას იძლევა რომ მიმდინარე პერიოდში ჩვენს ხელთ არსებული სწრაფი ეკონომიკური ცვლადების გამოყენებით მოხდეს მიმდინარე ეკონომიკური ზრდის შეფასება. ამ მიზნებისთვის გამოიყენება სხვადასხვა ტიპის მოდელები და მეთოდები. ბოლო პერიოდში მონაცემების მოპოვების წყაროების ზრდის, ასევე ზოგადად ახალი პროგნოზირების მეთოდების პარალელურად გაიზარდა ნაუქასთინგის მეთოდების რაოდენობაც. აღნიშნულ ნაშრომში განხილულია ნაუქასთინგის სხვადასხვა მიდგომა და მონაცემების გამოყენება, შესაბამისი ლიტერატურისა და მეთოდების შეფასება და ემპირიული ანალიზი საქართველოს მაგალითზე. ასევე, მოცემული ნაშრომი მიზნად ისახავს ტრადიციულ მოდელებთან ერთად მოხდეს თანამედროვე პროგნოზირების მეთოდების გამოყენებაც, როგორცაა მანქანური სწავლების ალგორითმები, და შემოწმდეს თუ რამდენად კარგადაა შესაძლებელი მათი მეშვეობით პროგნოზირება და თუ რამდენად უკეთესი ან/და უარესია ტრადიციულ მეთოდებთან შედარებით. მოცემული მეთოდები წარმოადგენს სიახლეს, რადგან გამოყენებულია არატრადიციული მონაცემები და მეთოდები მთლიანი შიდა პროდუქტის ნაუქასთინგისთვის.

2. ლიტერატურის მიმოხილვა

მაღალი სიხშირის მქონე ცვლადების გამოყენებას ეკონომიკური პროცესების, კერძოდ კი ბიზნეს ციკლების „მობრუნების“ წერტილების დროულად საპოვნელად დიდი ისტორია აქვს. ტრადიციული ნაუქასთინგის მეთოდები სწორედ ამ მიზნებისთვის გამოიყენებოდა, რომლებიც მოიცავდა მრავალი სწრაფი ინდიკატორის მეშვეობით, იმ ზოგადი არადაკვირვებადი ფაქტორების შეფასებას, რაც მიმდინარე ეკონომიკურ პროცესებს განსაზღვრავს. ეს იდეა პირველად გაჟღერდა (Sargent & Sims, 1977)-ის ნაშრომში, სადაც ისინი ცდილობენ მცირე რაოდენობის ფაქტორებით ბიზნეს ციკლის მერყეობების ახსნას. ამის შემდგომ (Stock & Watson, 1989), აჩვენეს რომ შესაძლებელია მრავალი მაკროეკონომიკური ცვლადების ერთობლივი მოძრაობა ასახული იყოს მცირე რაოდენობის ამ ცვლადებისათვის საერთო, ხელოვნურად შექმნილ „ლატენტურ“ ცვლადებში, რომელიც ეკონომიკური აქტივობისთვის თვალყურის დევნების საშუალებას იძლევა. სწორედ ეს მიდგომა, რომელსაც შემდგომ დინამიკური ფაქტორული მოდელი (Dynamic Factor Model – DFM) ეწოდა, შეიძლება ჩაითვალოს ტრადიციულ მეთოდად ეკონომიკური აქტივობის ნაუქასთინგის. DFM მოდელი თავისი შინაარსით წარმოადგენს N დაკვირვებადი ცვლადის გამოსახვას შემცირებული რაოდენობის დაუკვირვებადი ფაქტორებით და არაკორელირებული ნარჩენობითი წევრებით. DFM შეადგენს უფრო დიდი სიმრავლის ქვესიმრავლეს, რასაც ეწოდება state-space და „დამალული მარკოვის“ (Hidden Markov) მოდელები, სადაც დაკვირვებადი ცვლადები შეფასებული არიან დაუკვირვებადი ლატენტური ცვლადებით, რომლებიც თავად დროში ავტორეგრესიული პროცესით ან ვექტორული ავტორეგრესიით ვითარდებიან. დაკვირვებადი ცვლადის დამოკიდებულება ფაქტორებზე, შესაძლებელია წარმოდგენილი იყოს ორნაირად:

- როდესაც დამოკიდებულება დინამიკურია, ანუ t დროში დაკვირვებადი X ცვლადი შესაძლებელია დამოკიდებული იყოს t , $t-1$ და $t+n$ დროით პერიოდში მყოფ ფაქტორებზე, ამ მეთოდს ეწოდება DFM-ის დინამიკური ფორმა.
- ხოლო მეორე ფორმას წარმოადგენს სტატიკური (static, stacked) DFM, რის დროსაც როგორც დაკვირვებადი ისე დაუკვირვებადი ფაქტორები t დროში არიან წარმოდგენილნი.

DFM მოდელებში გამოყენებული ფაქტორების შეფასებამ განვითარების სამი ეტაპი გაიარა. **პირველი ეტაპის** ფაქტორების შეფასება ეყრდნობოდა მცირე განზომილების (სადაც ცვლადების რაოდენობა N მცირეა) პარამეტრულ მოდელებს, რომლებიც იხსნება გაუსის მაქსიმალური დამაჯერებლობის (Gaussian maximum likelihood) მეთოდით და კალმანის

ფილტრით ((Watson & Engle, 1983), Stock and Watson (1989), (Sargent, 1989) და (Quah & Sargent, 1993)). **მეორე თაობის** შეფასების მეთოდი წარმოადგენს არაპარამეტრულ შეფასების მეთოდს, როგორცაა წარმოადგენს კროს-სექციური საშუალო და Principal Component Analysis (PCA), სადაც გამოყენებული იყო დიდი რაოდენობის ცვლადები. ამ მეთოდის მთავარი შედეგი არის ის, რომ როდესაც დროითი მწკრივი და ცვლადების რაოდენობა საკმარისად დიდია, ფაქტორების შეფასების სიზუსტე იმდენად მაღალია, რომ შესაძლებელია მოდელში მათი რეგრესორებად გამოყენება. ხოლო **მესამე თაობის** შეფასების მეთოდი აერთიანებს პირველ და მეორე თაობის მიდგომებს. ამ მიდგომას ასევე უწოდებენ ორ საფეხურიან DFM-ს. პირველ ეტაპზე, არაპარამეტრული მეთოდით როგორცაა PCA, ფასდება ფაქტორები, რომლებიც შემდეგ გამოყენებულია კალმანის ფილტრში პირველადი კოეფიციენტების შესაფასებლად, პირველადი კოეფიციენტებით მიღებული ახალი ფაქტორები კვლავ გამოყენებულია მოდელში და ეს პროცესი გრძელდება იქამდე სანამ გამოთანაბრება ე.წ. კონვერგენცია არ მოხდება. ეს მეთოდი გამოყენებული და შემოთავაზებული იქნა (Giannone et al., 2008) და (Doz et al., 2006)- ის მიერ, სადაც მათ ეს მეთოდი გამოიყენეს კვარტალური მშპ-ს და ინფლაციის ნაუქასთისთვის აშშ ეკონომიკის მაგალითზე, რომელშიც გამოყენებული იყო 200 ამდე ყოველთვიური ცვლადი. ასევე DFM-ის გამოყენებით (Matheson, 2011) მოახდინა 32 განვითარებული და განვითარებადი ეკონომიკების რეალური კვარტალური მშპ-ს ნაუქასთინგი, რის მიხედვითაც უმეტეს შემთხვევაში DFM საუკეთესო შედეგით გამოირჩეოდა, ასევე ნაშრომში აღნიშნულია, რომ კარგი შედეგებით ხასიათდება მოდულების ნაუქასთების გაერთიანების (Pooling) მეთოდებიც.

DFM მეთოდი, ვინაიდან ის იყენებს კალმანის ფილტრს, საშუალებას იძლევა მიმდინარე ეკონომიკური აქტივობის მაღალი სიხშირის ინდექსი შეიქმნას, (Bańbura et al., 2013)-მა გამოიყვანეს სწრაფი ინდიკატორების გამოყენებით და დინამიკური ფაქტორული მოდელის მეშვეობით ყოველთვიური და ყოველდღიური მშპ-ს ზრდის ინდექსი, რომელიც მშპ-ს მერყეობის დიდ ნაწილს ხსნის. აღნიშნული მეთოდი, რაც მდგომარეობს მაღალი სიხშირის ეკონომიკური აქტივობის ინდექსის გამოყვანას, ასევე პოპულარული მეთოდია კვარტალური მშპ-ს ნაუქასთისთვის, ვინაიდან ის საშუალებას იძლევა ყოველ თვე მოხდეს მიმდინარე კვარტლის დინამიკის შეფასება.

(Bräuning & Koopman, 2014)-ის მიერ შემოთავაზებულ იქნა DFM-ის მოდიფიცირებული ვარიანტი - collapsed dynamic factor model (CDFM), რომელიც იყენებს DFM-ის უნარს შეფასებული იქნას დაუკვირვებადი ყოველთვიური მშპ.

(Bräuning & Koopman, 2014)-ის მეთოდი ფაქტორების შეფასებისათვის PCA-ის ნაცვლად ანალიზის ბოლო ეტაპზე, იყენებს Partial least squared (PLS) მეთოდს. PCA-სა და PLS-ს შორის განსხვავება

მდგომარეობს შემდეგში, რომ PCA ანალიზი მიეკუთვნება „დამოუკიდებელ“ (unsupervised) შეფასების მეთოდების რიგს, სადაც დამოკიდებული ცვლადი არ მონაწილეობს პარამეტრების შეფასებაში, ხოლო PLS ფაქტორების შეფასებისას ასევე უყრდნობა დამოკიდებულ ცვლადს, რაც საშუალებას იძლევა ფაქტორის ფორმირებისას უფრო მეტი წონა მიენიჭოს იმ ცვლადს რომელიც უფრო მეტად კორელირებს დამოკიდებულ ცვლადთან, განსხვავებით PCA მიდგომისა, სადაც წონები მხოლოდ ცვლადების ჯგუფის ვარიაციის ახსნადობის მიხედვით ნაწილდება.

მას შემდეგ რაც DFM-ის მიხედვით შევაფასებთ ყოველთვიური მშპ-ს ინდექსს, სწორედ ეს ინდექსი შეიძლება გამოყენებულ იქნას PLS-ში დამოკიდებულ ცვლადად, რათა მოხდეს ფაქტორების შეფასება კვარტალური მშპ-ს ნაუქასთისთვის. აღნიშნული მეთოდის გამოყენებით (Ginker & Suhoy, 2021)-მა მოახერხეს ისრაელის კვარტალური მშპ-ს ნაუქასთი და იპოვეს რომ მოცემული მეთოდი ზრდის ნაუქასთის სიზუსტეს.

Covid-19 პერიოდში DFM მეთოდოლოგიამ განსაკუთრებით აქტუალობა შეიძინა და იგი ხშირად გამოიყენებოდა მშპ-ს ნაუქასთისთვის. (Sampi Bravo & Jooste, 2020)-მა გამოიყენეს გუგლის მობილობის ინდიკატორები DFM მოდელში რათა შეეფასებინათ პანდემიის ეფექტი ინდუსტრიულ გამოშვებაზე ლათინურ ამერიკასა და კარიბის ყურისთვის. (Lewis et al., 2022)-მა კი შექმნეს ყოველკვირეული ეკონომიკური აქტივობის ინდექსი DFM-ის გამოყენებით. ასევე ამ საკითხთან დაკავშირებულ ნაშრომებს მიეკუთვნება (Baumeister & Guérin, 2020), (Chapman & Desai, 2021) და (Antolin-Diaz et al., 2021))

ასევე, ტრადიციულ ნაუქასთინგის მეთოდებს მიუკუთვნება მოდელები, რომელებიც აკავშირებენ მაღალი სიხშირის მონაცემებს დაბალი სიხშირის მონაცემთან. ყველაზე მარტივ ასეთ მეთოდს წარმოადგენს „გადამყვანი“ (Bridge) მოდელი, რომელმაც პოპულარობა მოიპოვა მისი სიმარტივისა და დაბალი ტექნიკური მოთხოვნილების მეშვეობით. აღნიშნულ მეთოდში გამოყენებულია ერთ განტოლებიანი რეგრესიები კვარტალურ მშპ-სა და მცირე რაოდენობის მაღალი სიხშირის მონაცემებზე, რომელიც აგრეგირებულია კვარტალურ სიხშირეზე სხვადასხვა მეთოდით - საშუალო არითმეტიკული, ჯამი ან რაიმე სხვა ტრანსფორმაცია (იხ (Baffigi et al., 2004)). იმ შემთხვევაში თუ ნაუქასთი ხდება კვარტლის მიმდინარეობისას ან იმ პერიოდში როცა ჯერ არაა გამოქვეყნებული ამხსნელი ცვლადები (“ragged”/“jagged” edge პრობლემა) ნაუქასთი/ პროგნოზი კეთდება შემდეგნაირად: ჯერ ხდება ამხსნელი ცვლადების პროგნოზირება ავტორეგრესიული მოდელების მიხედვით, რომელიც ხშირ შემთხვევაში მოითხოვს დიდი რაოდენობის ლაგებს, რომელიც თავის მხრივ ზრდის შესაფასებელი პარამეტრების რაოდენობას და შესაბამისად ამცირებს მოდელის თავისუფლების ხარისხსა და შედეგების სანდოობას. აღნიშნული პრობლემა გადაჭრილია mixed data sampling (MIDAS) მეთოდში, სადაც ლაგების

სტრუქტურა შეცვლილია არაწრფივი ფუნქციით (იხ. (Clements & Galvao, 2009), (Kuzin et al., 2011)). ამ მეთოდის ჯგუფს მიეკუთვნება შერეული სიხშირის ვექტორული ავტორეგრესიებიც (Mixed frequency VARs), რომელიც წარმოადგენს ერთ განტოლებიანი მოდელების ალტერნატივას (იხ. (Giannone et al., 2009) და (Kuzin et al., 2011)). ზემოთ აღწერილი მეთოდების ეფექტიანობა შეფასებულია (Camacho et al., 2013)-ის ნაშრომში.

იმის მიუხედავად, რომ ზემოთ მოყვანილი მოდელები წარმოადგენს ძირითად მეთოდს, რის მიხედვითაც სხვადასხვა ინსტიტუტებში ხდება ნაუქასთინგი, თანამედროვე პერიოდში სულ უფრო მეტ პოპულარობას იხვეჭს მანქანური სწავლების ალგორითმები მოცემული ამოცანის შესრულებისთვის. გარდა იმისა, რომ მანქანური სწავლების სხვადასხვა ალგორითმი განსხვავებული მეთოდით აფასებს პარამეტრებს, ასევე ის საშუალებას იძლევა მონაცემთა შორის ურთიერთკავშირის რთული არაწრფივი კავშირების პოვნისა და მონაცემებზე დაფუძნებული ნაუქასთის/პროგნოზის გაკეთებას, რაც იძლევა იმის პოტენციალს რომ შემცირდეს გადახრა შეფასებულ მნიშვნელობასა და რეალურ მონაცემს შორის.

მანქანური სწავლების გამოყენება გარდა ნაუქასთისა, ეკონომიკაში სხვადასხვა მიმართულებითაცაა შესაძლებელი. მაგალითად, მანქანური სწავლების მეთოდებით შესაძლებელია ცვლადების მნიშვნელოვნების დადგენა (Kohlscheen, 2021), პოლიტიკის ანალიზი ((Mariam Dundua & Otar Gorgodze, 2022),(Chakraborty & Joseph, 2017)) რეცესიის ალბათობის დადგენა (Basuchoudhary et al., 2020), პროგნოზირება (Ter-Martirosyan et al., 2018) და სხვა. იმის მიუხედავად რომ მანქანური სწავლების მეთოდები ძირითადად შემუშავებული და გამოყენებულია პროგნოზირებისთვის, მისი გამოყენება ნაუქასთისთვისაც იძენს აქტუალობას. (Richardson et al., 2021) იყენებს დაახლოებით 600 ადგილობრივ და საერთაშორისო ცვლადს ახალი ზელანდიის მშპ-ს პროგნოზირებისთვის, რის შედეგადაც ასკვნის რომ მანქანური სწავლების ალგორითმები ძირითადად კი **Boosted Trees, Support Vector Machine** და **Neural Networks** უკეთეს შედეგს იძლევიან ვიდრე მარტივი ავტორეგრესიული მოდელები და DFM.

სხვადასხვა მანქანური სწავლების მეთოდების გამოყენებით კვარტალურ მაკროეკონომიკურ და ფინანსურ მონაცემებზე (Muchisha et al., 2021) მსგავს დასკვნამდე მიდის, რომ მანქანური სწავლების მეთოდები (**Elastic Net, SuperLearner** და **Recurring Neural Network**) უკეთეს შედეგს იძლევა ვიდრე ავტორეგრესიული ბენჩმარკ მოდელი და სხვა ტრადიციული სტატისტიკური მეთოდები. ისევე როგორც DFM-ის შემთხვევაში, მანქანური სწავლების მეთოდებითაც შესაძლებელია ეკონომიკური აქტივობის ინდექსის გამოყვანა, (Woloszko, 2020)-მა გუგლის ტრენდების, სპეციფიკური სიტყვების ძიებისა და ნეირონული ქსელის გამოყენებით შექმნეს ეკონომიკური აქტივობის ყოველკვირეული ინდექსი OECD-ის ქვეყნებისათვის. მათი შედეგებით

მხოლოდ გუგლის ტრენდების და სპეციფიკური სიტყვების გამოყენებით მიღებული მოდელის შედეგები 17 პროცენტით უკეთესია ვიდრე AR(1) პროცესით მიღებული შედეგები.

(Dauphin et al., 2022)-ის ნაშრომში შესწავლილია DFM-ისა და მანქანური სწავლების მეთოდების ეფექტიანობა ნაუქასთინგისას კრიზისებსა და „ნორმალურ“ პერიოდში ევროპის ქვეყნებისთვის, მათი კვლევის მიხედვით DFM მეთოდი ნორმალურ პერიოდებში მაღალი სიზუსტით გამოირჩევა, მაშინ როცა მანქანური სწავლების მეთოდები ე.წ. მობრუნების წერტილების (როდესაც ბუმი იცვლება რეცესიით ან/და პირიქით) დადგენაში უკეთეს შედეგს იძლევიან. (Barhoumi et al., 2022) იყენებენ მანქანური სწავლების მეთოდებს სუბსაჰარული აფრიკის ქვეყნებისთვის, როგორც უკეთეს მეთოდს შეზღუდული მონაცემების პრობლემის დასაძლევად. მათი შედეგების მიხედვით მანქანური სწავლების მეთოდებით შესაძლებელია უკეთესი შედეგების მიღება, ვიდრე ეს ტრადიციული მეთოდებითაა შესაძლებელი.

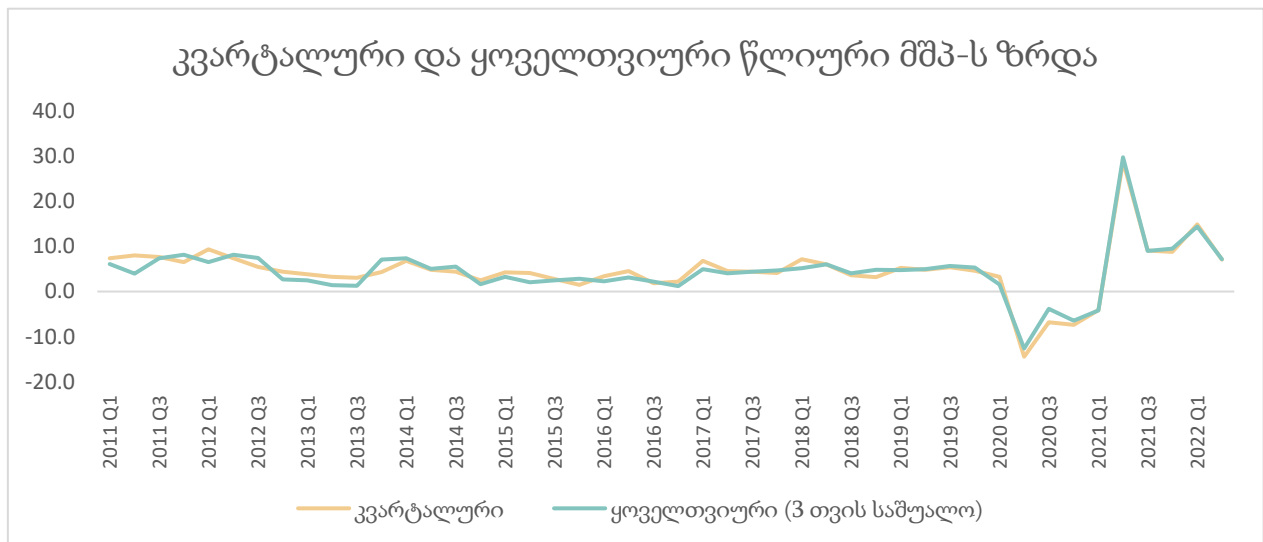
თავის ნაშრომში (Bolhuis & Rayner, 2020) თურქეთის მაგალითზე აჩვენებენ რომ მანქანური სწავლების მეთოდების და შემდგომ მათი შედეგების გაერთიანებით (Model Ensemble) ტრადიციულ მეთოდებთან შედარებით პროგნოზის და ნაუქასთის შეცდომის შემცირება საშუალოდ 30 პროცენტითაა შესაძლებელი. უფრო ადრინდელი ნაშრომები კი მოიცავს ლიბანის მშპ-ს ნაუქასთს (Tiffin, 2016), სადაც გამოყენებულია Elastic Net Regression (ENR) და Random Forest (RF) ალგორითმები. მათი შედეგების მიხედვით ENR უფრო უკეთეს შედეგს იძლევა ვიდრე RF, ხოლო მათი შედეგების გაერთიანებით (Model Ensemble) ENR-ის შედეგის გაუმჯობესება ხდება.

3. მონაცემები

3.1 მონაცემების თავისებურება საქართველოსთვის

ზემოთ მოყვანილი ლიტერატურის უმრავლესობა ახდენს კვარტალური მშპ-ს ნაუქასთინგს უფრო მაღალი სიხშირის მონაცემების დახმარებით სხვადასხვა მეთოდის გამოყენებით. აღნიშნული სიტუაცია საქართველოს მშპ-ს ნაუქასთინგის პრობლემისგან განსხვავებულია, რადგან საქართველოს სტატისტიკის ეროვნული სამსახური (საქსტატი) სხვა ქვეყნების მსგავსად (როგორცაა დიდი ბრიტანეთი, ყაზახეთი, სომხეთი , აზერბაიჯანი და სხვა) ყოველთვიურად აქვეყნებს წინა თვის მშპ-ს შეფასებას (იხ, გრაფიკი 1).

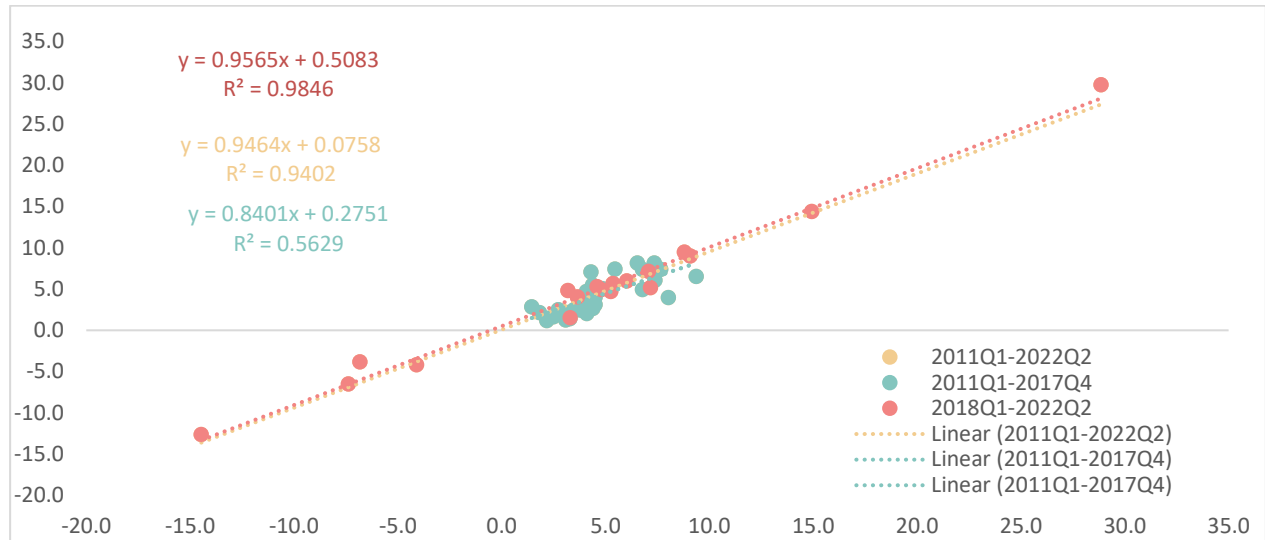
გრაფიკი 1. კვარტალური და ყოველთვიური წლიური მშპ-ს ზრდის ტემპი



წყარო : საქსტატი, ავტორის გაანგარიშებები

ეს საშუალებას იძლევა კვარტალური მშპ-ს ზრდის შესახებ ინფორმაცია მიღებული იყოს მხოლოდ ერთი თვის დაგვიანებით. საქსტატის მიერ ყოველთვიურად შეფასებული ეკონომიკური ზრდის მონაცემები მაღალი სიზუსტით იძლევა საშუალებას შეფასებული იყოს კვარტალური ზრდის ტემპიც, როგორც გრაფიკი 2-ზე ჩანს განსაკუთრებით ბოლო წლების განმავლობაში მაღალი სიზუსტით ხდება ეკონომიკური ზრდის შეფასება. ამიტომ, ტრადიციული მეთოდებისგან განსხვავებით, რომელთა მთავარ მიზანს წარმოადგენს კვარტალური ეკონომიკური ზრდის აწმყოში პროგნოზირება იმ დროისათვის არსებული მონაცემებით, საქართველოს შემთხვევაში ასევე მნიშვნელოვანია, რომ მოხდეს ყოველთვიური მშპ-ს მიმდინარე პერიოდში პროგნოზირება სწრაფი ეკონომიკური ინდიკატორებით. ამ ნაშრომის მიზანია ყოველთვიური ინდიკატორების გამოყენებით ყოველთვიური ეკონომიკური აქტივობის შეფასება.

გრაფიკი 2. კვარტალური და ყოველთვიური მშპ-ს ზრდის ტემპების ურთიერთკავშირი



წყარო: საქსტატი, ავტორის გამოთვლები

ვინაიდან ყოველთვიური ინდიკატორები, რომელიც თვის დასაწყისშივე ხელმისაწვდომია მცირე რაოდენობით მოიპოვება, ამიტომ საჭიროა სხვა არატრადიციული წყაროების და ცვლადების გამოყენება. ამისათვის ნაშრომში გამოყენებულია Google Trends³-ის მონაცემები.

3.2 მონაცემების აღწერა და დამუშავება

მოცემული კვლევისთვის გამოყენებული მონაცემები⁴ მოპოვებულია სხვადასხვა წყაროდან, რომლებსაც წარმოადგენს საქართველოს ფინანსთა სამინისტრო (Ministry of Finance - MOF), შემოსავლების სამსახური (Georgian Revenue Service – GRS), საქართველოს ეროვნული ბანკი (National Bank of Georgia - NBG), საქართველოს სტატისტიკის ეროვნული სამსახური (Geostat) და გუგლი (Google). აღნიშნული წყაროებიდან მონაცემები შესაძლებელია დაიყოს 3 ტიპად.

- პირველი წარმოადგენს მონაცემებს რომელთა შესახებ ინფორმაციაც ხელმისაწვდომია თვის დასრულებიდან ერთი კვირის განმავლობაში.
- მეორე ტიპის მონაცემები ხელმისაწვდომია ორი კვირის შემდგომ და
- მესამე ტიპის მონაცემებია, რომლებიც ქვეყნდება თვის მესამე-მეოთხე კვირაში.

მესამე ტიპის მონაცემებს მიეკუთვნება მონეტარული და ფინანსური სექტორის შესახებ მონაცემები (სესხები, დეპოზიტები, ფულადი აგრეგატები და სხვა), დეტალური ვაჭრობის შესახებ

³ ასევე მაღალი სიხშირის მონაცემებისთვის კარგ წყაროს წარმოადგენს Google Mobility-ის რეპორტი მაგრამ 2022 წლის ოქტომბრიდან Google წყვეტს Google mobility-ის რეპორტის საჯარო გამოქვეყნებას, ამიტომ ნაშრომში მოცემული მოდელები არ იყენებენ აღნიშნულ მონაცემებს.

⁴ მონაცემების ზუსტი აღწერა, წყარო და გამოქვეყნების თარიღი იხილეთ დანართ 8.1-ში

მონაცემები, სამრეწველო/იმპორტის/ექსპორტის ფასები, საბანკო ბარათებით განხორციელებული ოპერაციები, რეალური ეფექტური გაცვლითი კურსი. გარდა ოფიციალური სტატისტიკისა, თვის ბოლოს ქვეყნდება მონაცემები სხვადასხვა კვლევითი ორგანიზაციების მხრიდანაც, როგორცაა PMCG-ის შრომის ბაზრის კვლევა⁵(Employment Tracker), სადაც გასული თვის განმავლობაში ონლაინ დასაქმების პლატფორმაზე გამოქვეყნებული ვაკანსიების რაოდენობაა გაანალიზებული, ასევე ISET-PI-ის მიერ ქვეყნდება მომხმარებელთა ნდობის ინდექსი⁶ (Consumer Confidence Index), ყველა ეს ზემოთაღნიშნული მონაცემები თავისი შინაარსითა და ხარისხით კარგად აღწერენ ეკონომიკის მიმდინარე მდგომარეობას, თუმცა მათი გამოქვეყნება ხდება 3-4 კვირის დაგვიანებით, რაც ჩვენი მიზნებისთვის არაა ხელსაყრელი რადგან თვის ბოლო სამუშაო დღეს საქსტატის მიერ ქვეყნდება ეკონომიკური ზრდის წინასწარი შეფასება, ამიტომ დაგვიანებული მონაცემების გამოყენებას ნაუქასთისთვის მხოლოდ მცირედი ღირებულება გააჩნია. **მეორე ტიპის** მონაცემებში შედის ფულადი გადმორიცხვების შესახებ მონაცემები, ტურიზმიდან შემოსავლები, წინასწარი საქონლით საგარეო ვაჭრობის შესახებ მონაცემები და დამატებითი ღირებულების გადასახადის გადამხდელ საწარმოთა ბრუნვა, რომლებიც შემდგომი თვის 15-16 რიცხვისთვისაა ხელმისაწვდომი. **პირველი ტიპის** მონაცემებს კი მიეკუთვნება გაცვლითი კურსები, დროის რეალურ რეჟიმში ანგარიშსწორების სისტემაში (Real-time gross settlement - RTGS) განხორციელებული ოპერაციები, ლარის ფულადი ბაზრის მაჩვენებლები, თბილისის ბანკთაშორის საპროცენტო განაკვეთები, სამომხმარებლო ფასების ინფლაცია, საგადასახადო შემოსავლები, მთავრობის დანახარჯები, მონაცემები საქონლით საგარეო ვაჭრობაზე სასაზღვრო დეკლარაციებიდან და სალაროს ბრუნვის შესახებ მონაცემები ხელმისაწვდომია თვის დასრულებიდან პირველივე კვირაში.

როგორც ჩანს, პირველი ტიპის მონაცემების რაოდენობა, რომელთა შესახებ ინფორმაციაც ხელმისაწვდომია მცირედი დაყოვნებით, არც თუ ისე დიდია. ეს პრობლემა მხოლოდ საქართველოსთვის არაა დამახასიათებელი, ამიტომ სულ უფრო იზრდება ახალი გზებით სწორი შინაარსისა და ხარისხის მონაცემების მიღების საჭიროება. სწორედ ამ მიზნებისთვის სულ უფრო პოპულარული ხდება გუგლის მიერ მოწოდებული მონაცემები, როგორცაა გუგლ ტრენდი (Google Trends) და გუგლ მობილობა (Google Mobility), იმის მიუხედავად რომ ეს უკანასკნელი ერთ-ერთი საუკეთესო მაჩვენებელია ეკონომიკური აქტივობის საზომად, სამწუხაროდ 2022 წლის ოქტომბრიდან გუგლმა შეწყვიტა მისი განახლება და ის საჯაროდ ხელმისაწვდომი აღარაა. ხოლო რაც შეეხება Google Trends მონაცემებს, მისგან ბევრი სახის ინფორმაციის მიღება შეიძლება. ძირითადად Google Trends-იდან მიღებული მონაცემები ორი სახით გვხვდება

⁵ <https://pmcresearch.org/periodic/11>

⁶ <https://iset-pi.ge/en/indexes/5-consumer-confidence>

ლიტერატურაში, პირველი ესაა კონკრეტული სიტყვების (key words) ძიების სიხშირის შესახებ ინფორმაცია, (იხ. (Woloszko, 2020), (Narita & Yin, 2018)) ასევე როდესაც მონაცემები მიღებულია თემა/კატეგორიის შესახებ (იხ. (Barhoumi et al., 2022), (Robin, 2018), (Austin et al., 2021),(Carrière-Swallow & Labbé, 2010)).

სწორედ ეს უკანასკნელი მეთოდია არჩეული საქართველოსთვის Google Trends-ის ინდექსების გამოყენებისთვის, რადგან რომელიმე კონკრეტული სიტყვის ძიების ინდექსის გამოყენება შესაძლოა გამოსადეგი იყოს რაიმე მოვლენის წინასწარ განსაზღვრისთვის, მაგალითად რეცესიისთვის განვითარებულ, ძირითადად ლათინურ დამწერლობის მქონე ქვეყნებისთვის, რადგან რეცესიის პირობებში ისინი უფრო მიდრეკილნი არიან გუგლის საძიებო სისტემაში დაწერონ ისეთი სიტყვები რომელიც უკავშირდება სამსახურის ძიებას, უმუშევრობის ბენეფიტებს ვიდრე ეს არა ლათინური დამწერლობის მქონე ქვეყანაში, სადაც მსგავსი საკვანძო სიტყვებით რეცესიის ან რაიმე მსგავსი პროცესის პროგნოზირება რთულ პრობლემას წარმოადგენს. Google Trends ყოველთვიური კატეგორიების ინდექსები ხელმისაწვდომია თვის დასასრულიდან 72 საათის (3 დღის) შემდგომ რაც საშუალებას იძლევა ნაუქასთვის დამატებითი პირველი ტიპის მონაცემების იქნეს თავმოყრილი.

გარდა ხელმისაწვდომი ინფორმაციისა, ნაუქასთვის მონაცემების გამრავალფეროვნებისათვის, შესაძლებელია ე.წ. მონაცემების ინჟინერიის (Variable/Feature engineering) გამოყენებაც, რომლის საშუალებითაც შესაძლებელია ახალი მონაცემების შექმნა აღნიშნული მონაცემებისგან. მონაცემთა ინჟინერია შესაძლოა ბევრი გზით მოხდეს, მაგალითად შეიძლება ახალი ფიქტიური ცვლადების შექმნა, არსებული ცვლადების ტრანსფორმაცია, ცვლადების ერთმანეთზე გამრავლება/გაყოფა, ხარისხში აყვანა და ა.შ. ზოგ შემთხვევაში მონაცემთა ინჟინერია გამოიყენება ცვლადებს შორის არაწრფივი კავშირის წრფივი მოდელით შესაფასებლად. ამისათვის ხდება ცვლადის შესაბამის ხარისხში აყვანა, თუმცა ხშირად ახალი ცვლადების შექმნის უკან რაიმე თეორიული საწყისი არ არსებობს და ეს პროცესი ემსახურება წმინდა პროგნოზირების მოთხოვნებს. ჩვენს შემთხვევაში შესაძლებელია მონაცემთა ინჟინერია ისე, რომ ეს ორივე ნაწილი დაკმაყოფილებული იყოს.

ნაშრომში მოცემული მონაცემებისგან შესაძლებელია ახალი მონაცემების შექმნა, ისე რომ მას ჰქონდეს თეორიული საწყისი და შემდგომ შემოწმდეს თუ რამდენად გამოსადეგია ის მოდელირებისას. მაგალითად, ინფლაციისა და თბილისის ბანკთაშორისი საპროცენტო განაკვეთის სხვაობით შესაძლებელია ახალი ცვლადის შექმნა, რაც რეალური საპროცენტო განაკვეთი იქნება, ასევე გაცვლითი კურსის ვაჭრობის/ტურიზმის/ფულადი გადმორიცხვებზე გამრავლებით ვიღებთ მოცემული ცვლადების გამოსახულებას ლარებში, შემდგომ

შესაძლებელია ფასების ზრდის ტემპით მათი დეფლირება, რათა მივიღოთ რეალური ცვლადები. ასევე ფასებისა და საგადასახადო შემოსავლების/ დღგ-ს ბრუნვების/სახელმწიფო დანახარჯების ურთიერთკავშირით მიიღება ახალი ცვლადები, რომლებიც წარმოადგენს ამ უკანასკნელთა რეალურ მაჩვენებელს. საბოლოოდ, ნაშრომში მონაცემების ინჟინერია ხდება მხოლოდ იმ შემთხვევაში თუ ის ატარებს ეკონომიკურ შინაარსს, თუ ახლად შექმნილი ცვლადის კორელაცია უფრო მაღალია ვიდრე ორიგინალის მაშინ ის ანაცვლებს ამ ცვლადს მონაცემთა ბაზაში და პირიქით. ასევე, წრფივი მოდელების არაწრფივადობისთვის გარდაქმნილია მონაცემები მათი შესაბამის ხარისხში აყვანით, სადაც ლოგიკა იგივეა: თუ მისი კორელაცია მშპ-სთან აღემატება ორიგინალისას ის ანაცვლებს მას, სხვა შემთხვევაში არ ხდება მისი გამოყენება. ცვლადების არჩევანისთვის გამოიყენება ორმაგი ფილტრი, პირველი მონაცემის კორელაცია მშპ-სთან უნდა აღემატებოდეს 0.4-ს და ამასთან მისი სტანდარტული გადახრა მთლიანი მონაცემების საშუალო სტანდარტულ გადახრას 10%-ზე მეტად არ უნდა აღემატებოდეს

მონაცემთა ინჟინერიის და თავდაპირველ მონაცემთა გაფილტვრის შემდგომ პირველი ტიპის მონაცემების რაოდენობა 12 (თავდაპირველი ცვლადების რაოდენობა მონაცემთა ინჟინერიის და ფილტრის გარეშე შეადგენდა 41 ცვლადს შეადგენდა) ცვლადს შეადგენს, ხოლო პირველი და მეორე ტიპის მონაცემების რაოდენობა 17-ს შეადგენს (ფილტრამდე შეადგენდა 47 ცვლადს). შემდგომ მონაცემების ტრანსფორმაციისთვის გამოიყენება წლიური ზრდის ტემპები, რომლებიც შემდგომ სტრანდარტიზებულია ისე რომ მწკრივის საშუალო ნულის ტოლი ხოლო სტანდარტული გადახრა ერთის ტოლი იყოს, ამ უკანასკნელს განსაკუთრებული მნიშვნელობა აქვს PCA, PLS, DMF და მანქანური სწავლების მეთოდებისთვის, რათა არ მოხდეს აბსოლუტურ მაჩვენებელში მაღალი ვარიაციის მქონე ცვლადებისთვის მეტი წონის მინიჭება, მაშინ როდესაც ისინი არ არიან უკეთესი ამხსნელი ცვლადები.

მონაცემები მოიცავს 01.2012-დან - 09.2022-მდე პერიოდს. კვლევაში თავდაპირველად გამოყენებულია მხოლოდ ის ცვლადები, რომლებიც თვის პირველ კვირაშია ხელმისაწვდომი, ხოლო შემდგომ კი შეფასებულია თუ რამდენად ზრდის ნაუქასთინგის უნარს მეორე ტიპის მონაცემების დამატება.

4. მოდელები და მეთოდოლოგია

პუბლიკაციაში ყოველთვიური ეკონომიკური აქტივობის ნაუქასთისთვის გამოყენებულია როგორც ტრადიციული მეთოდები - როგორცაა DFM ასევე, მანქანური სწავლების მეთოდები.

4.1 DFM მეთოდოლოგია

გამოყენებული DFM მოდელი წარმოადგენს მესამე თაობის DFM მოდელს სტატიკურ ფორმაში, რომელსაც შემდეგი სახე აქვს:

$$X_t = \Lambda f_t + e_t$$

$$f_t = \psi(L)f_{t-n} + \eta_t$$

სადაც $X_t = (X_{1t}, X_{2t}, \dots, X_{nt})$ წარმოადგენს $(n \times 1)$ სტაციონალურ პროცესს, სადაც n არის ცვლადების რაოდენობა ხოლო t კი დაკვირვებების რიცხვი. Λ არის $(n \times r)$ მატრიცა, რომელიც f_t ფაქტორების პარამეტრებს (Factor loadings) შეიცავს. ხოლო e_t არის იდიოსინკრატიული შოკი, რომელიც არ კორელირებს ფაქტორებთან, და მისი მათემატიკური ლოდინი (საშუალო) ნულის ტოლია. f_t ფაქტორების ურთიერთდამოკიდებულება მოდელირებულია, როგორც p რიგის VAR პროცესი. მოცემული განტოლებების შეფასება ხდება ორ ეტაპად, პირველ ეტაპზე ფაქტორების კოეფიციენტების (Factor loading) მოსაპოვებლად გამოიყენება PCA ანალიზით მიღებული კომპონენტების პარამეტრები. მეორე ეტაპზე ხდება მოცემული პარამეტრების პირობებში ახალი ფაქტორების შეფასება კალმანის ფილტრის მეშვეობით. პირველ ეტაპზე პარამეტრების წინასწარი შეფასებისთვის გამოყენებულია PCA ანალიზი და არ ხდება CFDM-მოდელის გამოყენება PLS⁷ მეთოდოლოგიით შემდგომი მიზნების გამო: ვინაიდან PLS-ის მეშვეობით ფაქტორების შეფასებისთვის საჭიროა დამოკიდებული ცვლადი (ამ შემთხვევაში მშპ-ს ზრდის ტემპი), შეუძლებელია მიმდინარე პერიოდში PLS მოდელით მოხდეს კომპონენტების მიღება, რომელიც შემდგომ DFM-ში იქნება გამოყენებული, რადგან საჭიროა ასახსნელი ცვლადის (ამ შემთხვევაში მშპ-ს ზრდის ყოველთვიური მონაცემი) t დროში, ამისათვის კი საჭიროა უფრო დაბალი სიხშირის მშპ-ს ინდექსის შექმნა, რაც სცდება აღნიშნული ნაშრომის მიზნებს. სტატიკური ფაქტორების რაოდენობის დასადგენად გამოიყენება Scree plots-ის ვიზუალური დიაგნოსტიკა. პირველ ეტაპზე PCA ანალიზისთვის შესაძლებელია გამოყენებული იყოს ორი

⁷ კვლევის მიმდინარეობისას მოხდა კომპონენტების შეფასება PLS მეთოდოლოგიითაც, რომლის მიხედვითაც უკეთესად ხდება მშპ-ს ახსნა. ვინაიდან ის წარმოადგენს უკეთეს კომპონენტის მიღების მეთოდს, მისი გამოყენება შეიძლება ფაქტორულ ვექტორულ ავტორეგრესიებში (Factor Augmented VAR – FAVAR), სადაც არ არის საჭირო ფაქტორების მიმდინარე t დროში არსებობა.

მიდგომა - „ნაივი“ (Naïve) PCA და სტრუქტურული PCA. პირველი მიდგომის თანახმად კომპონენტების ამოღება ხდება მთლიანი ცვლადების ჯგუფიდან და კომპონენტების რაოდენობა განისაზღვრება ისე რომ მიღებული კომპონენტების მიერ მთლიანი ვარიაციის 60-80% იყოს ახსნილი, ხოლო მეორე მეთოდის მიხედვით ხდება თავდაპირველი სექტორების იდენტიფიცირება (მაგალითად ფინანსური, რეალური ან სხვა საერთო მახასიათებლებით დაკავშირებული ცვლადები) რის შემდეგაც ხდება თითოეული სტრუქტურულად იდენტიფიცირებული ჯგუფიდან ერთი კომპონენტის ამოღება შემდგომ მოდელში გამოსაყენებლად. ნაშრომში გამოყენებულია PCA-ს პირველი მეთოდი. ფაქტორების VAR მოდელის ლაგების რაოდენობა განსაზღვრულია მოდელის გარეთ აკაიკ ინფორმაციულ კრიტერიუმზე დაყრდნობით (Akaike information criterion – AIC). ფაქტორების მიღების შემდგომ ყოველთვიური მშპ-ს ნაუქასთი ხდება სტატისტიკური უმცირეს კვადრატთა მეთოდით

$$Y_t = \alpha + \beta_1 F_t^1 + \beta_2 F_t^2 + \dots + \beta_n F_t^n$$

4.2 მანქანური სწავლების მეთოდები და მოდელები

ნაშრომში მანქანური სწავლებისთვის გამოყენებულია შემდეგი ალგორითმები - Regularized regression (Ridge, Lasso, Elastic Net) , Support Vector Machine – SVM, Decision Trees, Random Forest, Gradient Boosting (Ada, XGboost, LGBM), Neural Networks (NN) და Model Ensemble (Stacking, Blending).

ცხრილი 1. მანქანური სწავლების მეთოდების მოკლე აღწერა⁸

მეთოდი	აღწერა
რეგულარიზებული რეგრესიები - Lasso, Ridge, Elastic Net	LASSO (least absolute shrinkage and selection operator), Ridge რეგრესია და Elastic Net წარმოადგენენ წრფივი რეგრესიის მოდიფიცირებულ ვარიანტებს, სადაც სხვადასხვა ტიპის რეგულარიზაციაა შემოტანილი, რათა მოხდეს მოდელის პროგნოზირების უნარის გაუმჯობესება. ტრადიციულ რეგრესიის მოდელებთან შედარებით, აღნიშნული მოდელები კარგად ართმევენ თავს დიდი რაოდენობით ცვლადების მოდელში შეტანას, მულტიკოლინიარობას და მოდელის ზედმეტად მორგებას (overfitting) მაგრამ ამ ყველაფრის მიუხედავად ისინი არაწრფივადობას თავს ვერ ართმევენ რადგან ისინი ამ ყველაფრის გათვალისწინებით მაინც წრფივ მოდელებს წარმოადგენენ.

⁸ მანქანური სწავლების მოდელების დეტალური აღწერისთვის იხილეთ დანართი 8.2

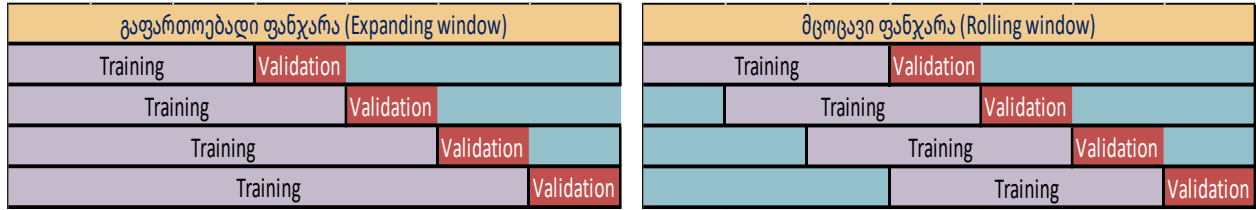
კერნელ მეთოდები - Support Vector Machine (SVM)	SVM მოდელი ქმნის ე.წ. ჰიპერპლანებს (hyperplanes) რათა დაეყოს ამხსნელი ცვლადების კომბინაციები და შემდგომ აკეთებს წერტილოვან პროგნოზებს თითოეული სექციისთვის, როგორც ეს კერნელურ (kernel) რეგულარიზირებულ რეგრესიებში ხდება. SVM თავს კარგად ართმევს ისეთ პრობლემებს რომელიც დაკავშირებულია არაწრფივადობასთან, მულტიკოლინიარობასთან, დიდი რაოდენობის ცვლადების მოდელში შეტანასთან და overfitting-თან, მაგრამ მისი შეფასების უნარი დამოკიდებულია კერნელის ფუნქციის და რეგულარიზაციის პარამეტრების სწორ შერჩევაზე.
გადაწყვეტილების ხეები - Decision Trees (DT)	გადაწყვეტილების ხეები არის დიაგრამის მსგავსი სტრუქტურები, რომლებიც შექმნილია კონკრეტული გამოსავლის პროგნოზირებისთვის. მაგალითად ავიღოთ დიაგრამა, სადაც თითოეული დონე არის კითხვა ორობითი (binary) დიახ ან არა პასუხით (მაგალითად, „დღეს გადამხდელ საწარმოთა ბრუნვის ზრდა 20%-ზე მაღალია?“) რასაც მოჰყვება სხვა დონეები ორობითი პასუხებით. დიაგრამის მიდევნება და კითხვებზე სათითაოდ პასუხის გაცემა, საბოლოოდ იძლევა საწყისი პრობლემის გადაჭრას. გადაწყვეტილების ხეები კარგად მუშაობს პრობლემებისთვის, სადაც მნიშვნელოვანია არაწრფივი ურთიერთქმედებები. თუმცა, პოტენციური მინუსი არის ის, რომ მეთოდი ძალიან მოქნილია და ამიტომ მიდრეკილია overfitting-ისკენ - საკმარისი კითხვებით, დიდ ხეს შეუძლია დააღაგოს თითოეული დაკვირვება საკუთარ ინდივიდუალურ უჯრაში (bin), რაც შესანიშნავად მოერგება ასახსნელი ცვლადის თითოეულ დაკვირვებას.
Random Forest - RF	RF წარმოადგენს პროგნოზების კომბინაციებს, რომლებიც მიღებულია ინდივიდუალური რეგრესიის ხეებისგან (Regression Trees). როგორც არაპარამეტრული ალგორითმი, RF კარგად ართმევს თავს წრფივი რეგრესიის პრობლემებს, რომელიც დაკავშირებულია არაწრფივადობასთან, მულტიკოლინიარობასთან და დიდი რაოდენობის ცვლადების მოდელში შეტანასთან. მიუხედავად მისი მოქნილობისა RF-ის უნარი მოახდინოს outlier-ების პროგნოზირება შეზღუდულია.
Gradient Boosting- (XGboost, Catboost) Adaptive boosting - AdaBoost	Gradient boosting ალგორითმი მსგავსია RF ალგორითმისა, რადგან ის გულისხმობს გადაწყვეტილების ხეების დიდი რაოდენობის გაერთიანებას. მაგრამ იმის მაგივრად, რომ მოხდეს ყველა მოდელის საშუალო გამოყენება, მოცემული მეთოდი ამას თანმიმდევრულად აკეთებს, სადაც თითოეული გადაწყვეტილების ხე აგებულია სათითაოდ, სადაც ყოველი ხე ცდილობს წინა ხის „შეცდომები“ გამოასწოროს. პირველი ხის პროგნოზებზე დაყრდნობით, მეორე ხე ცდილობს შეაფასოს პირველი ხის შეცდომები, შემდგომ მესამე ხე იმეორებს იგივეს მეორე ხისთვის და ა.შ. საბოლოო პროგნოზი არის ყველა ინდივიდუალური ხის პროგნოზის ჯამი. Adaptive boosting ოდნავ განსხვავებულ მიდგომას იყენებს. წინა მოდელის შეცდომების პროგნოზირების ნაცვლად, ყოველი გამეორება ცდილობს პროგნოზების გენერირებას ხელახლა შეწონილი (reweighted) მონაცემთა ბაზის საფუძველზე, სადაც წონები განისაზღვრება წინა მოდელით. მეტი წონა ენიჭება იმ შემთხვევებს, რომელსაც წინა მოდელი ცუდად გაუმკლავდა, და ნაკლები წონა ენიჭება შემთხვევებს, რომლებსაც წინა მოდელმა კარგად გაართვა თავი. საბოლოო პროგნოზი არის ყველა მოდელის შეწონილი ჯამი, წონებით განისაზღვრება თითოეული მოდელის სიზუსტით
ნეირონული ქსელები - Neural Networks (NN)	NN წარმოადგენს მრავალშრიან არაწრფივ მეთოდს, რომელიც აკავშირებს რიგ მონაცემებს ასახსნელ ცვლადთან. ქსელი შედგება შრეებისგან, რომლებიც წარმოადგენენ ხელოვნურ ნეირონებს (ასევე ხშირად მოიხსენიება როგორც კვანძი - node). თითოეული ნეირონი იღებს მონაცემს წინა შრის ნეირონებიდან, შემდეგ ტრანსფორმაციისთვის იყენებს ფუნქციას რათა მიღებული მონაცემები გარდაქმნას ერთ შედეგად (output), და შემდგომ ამ შედეგს აგზავნის შემდეგ შრეში. NN არის ძალიან მოქნილი, ვინაიდან შესაძლებელია სხვადასხვა ფუნქციის გამოყენება თითოეულ ნეირონში, ასევე შესაძლებელია თითოეული შრის სხვადასხვა სტრუქტურით იდენტიფიცირება NN როგორც დახვეწილი და მოქნილი ალგორითმი, მიჩნეულია როგორც მძლავრი ინსტრუმენტი პროგნოზირებისთვის, რომელიც ავარებს ტრადიციული რეგრესიის მეთოდების ნაკლოვანებებს.
შედეგების გაერთიანება - Stacked generalization	Stacking წარმოადგენს feedforward ნეირონულ ქსელს, რომელიც იყენებს მოდელის დაჯგუფების (Wolpert, 1992)-ის მეთოდს. ამ მეთოდის მიხედვით ყოველი მოცემული მონაცემისთვის, ტრადიციული ნეირონული ქსელის

	<p>ტრანსფორმაციის ფუნქციისგან განსხვავებით, რომელიც იყენებს სიგმიოდურ ან სხვა რაიმე ფუნქციას, გამოიყენება ნებისმიერი მანქანური სწავლების ალგორითმი. რომლისგან მიღებული შედეგიც გადაეცემა შემდეგ შრეს, სადაც ასევე მანქანური სწავლების მეთოდებით ხდება მისი ტრანსფორმაცია.</p>
--	---

4.3 მოდელების შერჩევა და ნაუქასთინგის სიზუსტის შეფასება

მანქანური სწავლების მოდელების შეფასება რამდენიმე ეტაპად ხდება, სადაც საჭიროა გათვალისწინებული იყოს რამდენიმე ელემენტი. მანქანური სწავლების მოდელებისათვის საჭიროა ზოგიერთი პარამეტრის (ჰიპერპარამეტრის) მოდელის გარედან მიწოდება, რათა შემდგომ ალგორითმმა კარგად გაართვას თავი დავალებას. სწორი ჰიპერპარამეტრების განსაზღვრა მნიშვნელოვანი ეტაპია მანქანურ სწავლებაში და საბოლოო შედეგი დიდ წილადაა მასზე დამოკიდებული. ამიტომ მოდელის შეფასების საწყის ეტაპზე საჭიროა იმ ოპტიმალური (ან სუბოპტიმალური რადგან ხშირად შეუძლებელიცაა ნამდვილად ოპტიმალური ჰიპერპარამეტრების წყვილის დადგენა) ჰიპერპარამეტრების განსაზღვრა რასაც შემდგომ გამოვიყენებთ მთლიანი მოდელის შეფასებისას. ამისათვის კი საჭიროა საწყისი მონაცემები დაიყოს 2 ნაწილად - „სავარჯიშო „ (Training sample) და სატესტო მონაცემებთა ნიმუშებად (Test sample). სატესტო მონაცემების ნიმუშები დაახლოებით 10-15 პროცენტს უნდა შეადგენდეს. სავარჯიშო ნიმუშის ფარგლებში კი ხდება მოდელის ჰიპერპარამეტრების დადგენის პროცესი. სავარჯიშო ნიმუშში ხდება ე.წ კროს-ვალიდაცია, რის პროცესშიც ხდება ფსევდო out-of-sample პროგნოზების გაკეთება, რომელიც რამდენიმე ეტაპად ხორციელდება (k-fold cross-validation). დროითი მწკრივებისთვის არსებობს ორი კროს-ვალიდაციის მეთოდი - მცოცავი ფანჯარა (Rolling window) და გაფართოებადი ფანჯარა (Expanding window) (იხ. გრაფიკი 3). მეთოდის შერჩევა დამოკიდებულია იმაზე, თუ რომელი უფრო ახლოსაა პრაქტიკაში შესასრულებელ სამუშაოსთან. თუ ხშირად ხდება წარსული მონაცემების განახლება ან თუ ხშირია წარსული მონაცემების აქტუალობის დაკარგვა, მაშინ ამ ეტაპზე უმჯობესია შერჩეული იყოს მცოცავი ფანჯრის მეთოდი. თუ სამუშაოს დროს ყოველ ჯერზე როცა მონაცემების განახლება ხდება, დაკვირვება ემატება მონაცემთა ბაზას და შემდგომ ახლიდან ხდება მოდელის შეფასება და პროგნოზი, მაშინ უკეთესია გაფართოებადი ფანჯრის მეთოდი. სწორედ ეს უკანასკნელი ასახავს უფრო სწორად საქართველოსთვის ნაუქასთინგის პროცესს, ამიტომ კროს ვალიდაციისთვის სწორედ ეს მეთოდი იქნა შერჩეული.

გრაფიკი 3. დროითი მწკრივებისთვის კროს-ვალიდაციის მეთოდები



მას შემდგომ რაც კროს-ვალიდაციის მეთოდით მოხდება ოპტიმალური ჰიპერპარამეტრების დადგენა, ხდება მოდელების მთლიან სავარჯიშო ნიმუშზე შეფასება და სიზუსტის შემოწმება სატესტო ნიმუშზე. მსგავსი მეთოდი გავრცელებულია პროგნოზირებაში შეჯიბრებისას, ისეთ პლატფორმებზე, როგორიცაა კაგლი (Kaggle⁹). იმის მიუხედავად რომ მსგავს შეჯიბრებებში აღნიშნული რუტინა ეფექტიანია, პრაქტიკაში ის ზოგ ნაკლოვანებებს აჩენს, რადგან შესაძლებელია ამ პროცესისას მოხდეს მოდელების გაუმჯობესება სატესტო ნიმუშის გავლენით, რადგან ხდება მოდელების მოდიფიკაცია იქამდე სანამ ის უკეთეს შედეგს არ აჩვენებს სატესტო ნიმუშზე, ამ მოვლენას სატესტო ნიმუშის გაჟონვა ეწოდება (Test sample leakage). ამიტომ კვლევაში ამ მეთოდთან ერთად გამოყენებულია სხვა მეთოდიც, რაც ბექტესტინგის სახელითაა (Back testing) ცნობილი. ის საშუალებას იძლევა მოხდეს ფსევდო out-of-sample პროგნოზების გაკეთება. მაგალითად 2015 წლის იანვრიდან დაწყებული მოხდება მოდელის შეფასება 2015 წლის იანვრისთვის ხელმისაწვდომი მონაცემებით და შემდგომ გაკეთდება ნაუქასთი, ამის შემდგომ დაემატება შემდგომი თვის მონაცემები და ნაუქასთი გაკეთდება შემდეგი თვისთვის და ა.შ, საბოლოოდ რომელ მოდელსაც ჯამში უფრო მცირე შეცდომა ექნება ბექტესტისას ის იქნება საუკეთესო მოდელი. სიზუსტის შეფასება მოხდება ორი ინდიკატორით. პირველი ესაა Root Mean Squared Error (RMSE) ინდიკატორი.

$$RMSE = \sqrt{\frac{\sum_{t=1}^N (y_t - \hat{y}_t)^2}{N}}$$

რომელიც აფასებს საშუალო შეცდომის სიდიდეს ნაუქასთსა (\hat{y}) და რეალიზებულ (y) მნიშვნელობას შორის. მხოლოდ ამ ინდიკატორით შეფასებისას შეიძლება არასწორ დასკვნებამდე მივიდეთ, რადგან ნაუქასთის სიზუსტის სიდიდესთან ერთად მნიშვნელოვანია მან ასევე სწორი მიმართულებაც გვაჩვენოს, რაც გულისხმობს ზრდა/კლებას წინა თვის მონაცემთან შედარებით. ამ მიზნებისათვის გამოიყენება Mean Directional Accuracy (MDA) ინდიკატორი.

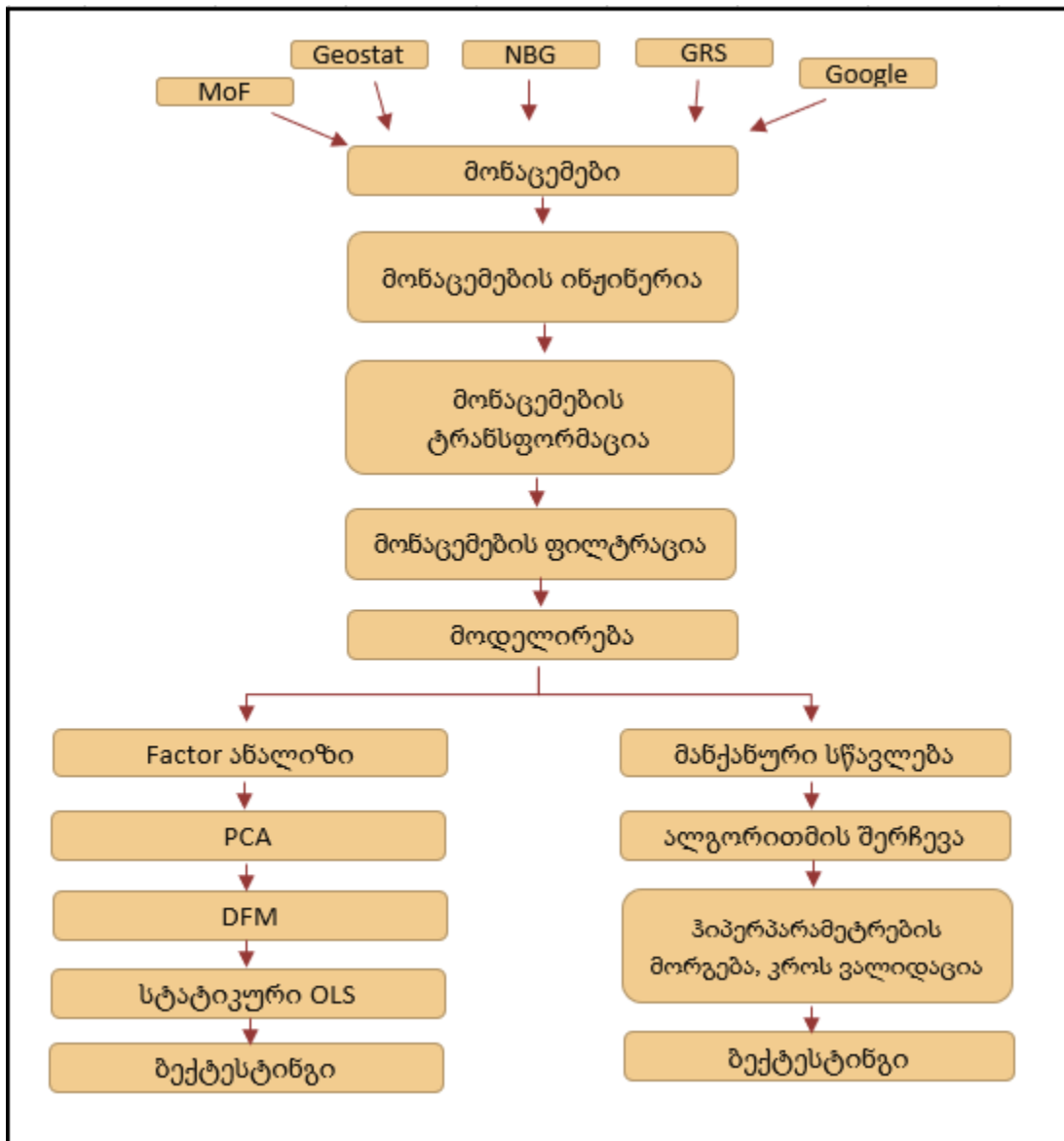
⁹ www.kaggle.com

$$MDA = \frac{1}{N} \sum 1[\text{sgn}(y_t - y_{t-1}) = \text{sgn}(\hat{y}_t - y_{t-1})]$$

$$\text{სადაც } 1_A = \begin{cases} 1 & \text{თუ } x \in A \\ 0 & \text{თუ სხვა} \end{cases} \text{ და } \text{sgn}(x) = \begin{cases} -1 & \text{თუ } x < 0, \\ 0 & \text{თუ } x = 0, \\ 1 & \text{თუ } x > 0 \end{cases}$$

ნაშრომში გამოყენებული მოდელირების პროცესი შემდეგნაირად გამოიყურება (გრაფიკი 4):

გრაფიკი 4. მოდელირების პროცესი



მოდელების შერჩევის და ოპტიმიზაციის პროცესში გამოყენებულია სხვადასხვა სახის ავტომატური მანქანური სწავლების და ხელოვნური ინტელექტის პაკეტი, რისი მეშვეობითაც შეფასდა ძალიან დიდი რაოდენობის მოდელები და მათი პარამეტრები, რათა საბოლოოდ ყოველი მანქანური მოდელის კლასიდან ნაპოვნი ყოფილიყო საუკეთესო მოდელი თავისი პარამეტრებით.

5. შედეგები

მიღებული შედეგები (ცხრილი 2) შერეული ხასიათისაა. პირველ რიგში უნდა აღინიშნოს რომ მოცემული მოდელების უმრავლესობა მშპ-ს ზრდის ტრაექტორიას მაღალი სიზუსტით პროგნოზირებს, მოდელების უმეტეს შემთხვევაში სიზუსტის მაჩვენებელი 70%-ს აღემატება. ასევე თვალსაჩინოა რომ მეორე ტიპის მონაცემების დამატებით გაცილებით მეტი სიზუსტის მიღწევას შესაძლებელი, საშუალოდ მოდელის შეცდომა მეორე ტიპის მონაცემების გამოყენების შემდგომ 20%-ით მცირდება.

ცხრილი 2. მიღებული მოდელების შედეგები.

მოდელები	პირველი ტიპის მონაცემები		მეორე ტიპის მონაცემები	
	RMSE	MDA	RMSE	MDA
AR(1)	5.71	66.2%	5.71	66.2%
DFM eviws in sample	4.45	70.1%	3.40	71.6%
LASSO	3.26	70.6%	2.58	76.5%
Ridge	3.25	69.1%	2.55	76.5%
Elastic Net	5.04	63.2%	3.33	67.6%
Xgboost	6.04	64.7%	5.28	70.6%
Catboost	5.38	66.2%	4.56	73.5%
SVR	3.21	69.1%	2.48	73.5%
Linear SVR	3.42	69.1%	2.51	70.6%
Random Forest	5.94	63.2%	5.23	70.6%
Adaboost	6.31	63.2%	5.68	64.7%
Neural Network	3.74	61.8%	2.58	66.2%
Stack Generalization	4.98	69.1%	3.34	75.0%
მოდელების საშუალო	4.00	69.1%	3.00	57.4%

წყარო : ავტორის გამოთვლები

ანალიზმა აჩვენა, რომ პირველი ტიპის მონაცემების გამოყენებისას უმეტესი მოდელები უკეთეს შედეგს იძლევიან ბენჩმარკ AR(1) მოდელთან შედარებით. AR(1) მოდელთან შედარებით უარესი შედეგის მქონე მოდელები ძირითადად წარმოადგენენ რეგრესიის ხესთან დაკავშირებულ მოდელებს, როგორებიც არიან Random Forest, Adaboost, Catboost, Xgboost. აღნიშნული შედეგი

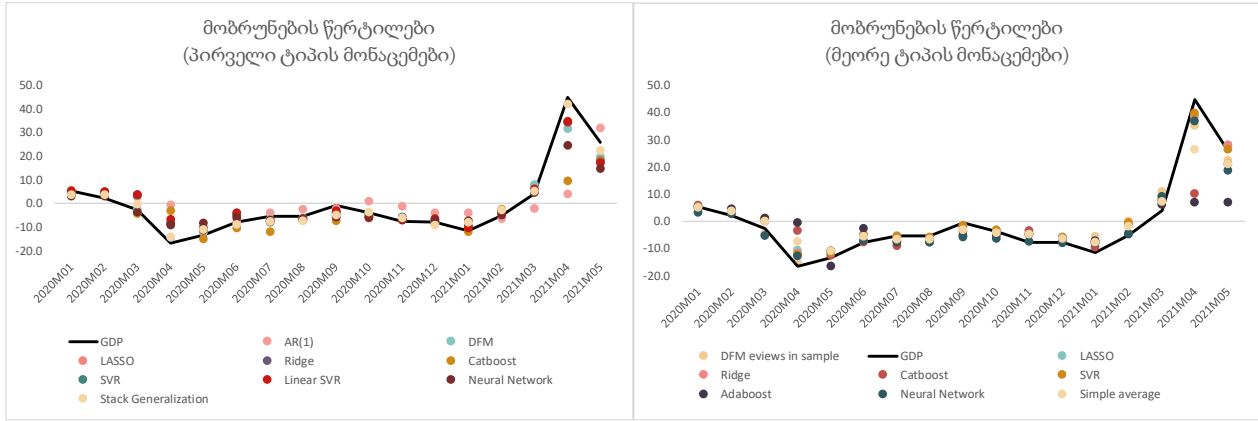
მოსალოდნელიც იყო, რადგან ანალიზი მოიცავდა, როგორც მაღალი ისე დაბალი მერყეობის პერიოდებს, ხოლო მსგავსი ტიპის მოდელები არაპარამეტრული ხასიათისაა, რაც საშუალებას არ იძლევა მოხდეს ისეთი პერიოდების მაღალი სიზუსტით შეფასება, რომელიც მოდელის შეფასებისას არ ფუნქციონირებდა. მსგავსი ტიპის მოდელების გამოყენება მიზანშეწონილია დაბალი მერყეობის პერიოდებისათვის. აღნიშნული შედეგები ვრცელდება მეორე ტიპის მონაცემების დამატების შემდგომაც.

პირველი ტიპის მონაცემებისათვის DFM მეთოდთან შედარებით უკეთეს შედეგებს იძლევა მანქანური სწავლების მეთოდები - Lasso, Ridge, SVR, Linear SVM, Neural Network. ყველაზე უკეთესი შედეგით გამოირჩევა Linear SVM, რომელიც DFM-თან შედარებით 27.8 პროცენტით უკეთეს შედეგს იძლევა. საშუალოდ კი პირველი ტიპის მონაცემებისთვის მანქანური სწავლების მეთოდები (გარდა გადაწყვეტილების ხის ტიპის მოდელებისა) DFM-თან შედარებით 17.9 პროცენტით უკეთეს შედეგს იძლევა. მიღებული შედეგები თანხვედრაშია გავრცელებულ ლიტერატურასთან, სადაც ნაუქასთინგისას რეგულარიზებული ტიპის მოდელები ხშირ შემთხვევაში უკეთეს შედეგს იძლევიან, ვიდრე სხვა უფრო კომპლექსური მოდელები.

რაც შეეხება მეორე ტიპის მონაცემებს აქაც მსგავსი სურათია, ერთის გამოკლებით როდესაც Elastic Net მეთოდოლოგიამ ნახტომისებური გაუმჯობესება განიცადა მეორე ტიპის მონაცემების დამატების შემდგომ. აღსანიშნავია NN მოდელის დაბალი შეცდომის მაჩვენებელი, იმ პირობებში რომ აღნიშნული ტიპის მოდელები ძირითადად მაღალი სიზუსტით გამოირჩევიან დიდი რაოდენობის მონაცემების არსებობისას. რეგულარიზებული წრფივი მოდელები, კერძო ტიპის მოდელები და ნეირონული ქსელის ტიპის მოდელები უკეთეს შედეგებს იძლევიან ვიდრე ტრადიციული მეთოდები, ეს შედეგები უცვლელია მონაცემების ტიპის მიუხედავად. აღნიშნულ შემთხვევაშიც საუკეთესო შედეგით გამოირჩევა Linear SVM, რომლის შედეგიც 26.2 პროცენტით აღემატება DFM-ს. ხოლო მანქანური სწავლების საშუალო მაჩვენებელი (გარდა გადაწყვეტილების ხის ტიპის მოდელებისა) 21.4 პროცენტით უკეთეს შედეგს აჩვენებს.

მოდელების შედეგების გაერთიანებით (Stacked Generalization), ორივე ტიპის მონაცემების პირობებში მიღებული შედეგები უკეთეს შედეგს ვერ აჩვენებენ მარტივ საშუალოსთან შედარებით. პირველი ტიპის მონაცემებისთვის DFM მოდელით უკეთესი შედეგი მიიღწევა ვიდრე პროგნოზების გაერთიანებით, ხოლო მეორე ტიპის მონაცემებისთვის DFM-თან შედარებით მცირედი გაუმჯობესება შეინიშნება. მოცემული ანალიზის პირობებში, მოდელების შედეგები გაერთიანებულია წრფივი რეგრესიის მეშვეობით, რადგან წრფივი რეგრესიის პირობებში გაერთიანებული შედეგები საუკეთესო შედეგით გამოირჩეოდა სხვა ალგორითმებთან შედარებით.

გრაფიკი 5. მობრუნების წერტილები



წყარო: ავტორის განგარიშებები

აღსანიშნავია, რომ მობრუნების წერტილების „დაჭერა“ კარგად ხდება უმეტესი მოდელების მიერ, რაც ერთის მხრივ ხაზს უსვამს მოდელების სწორად იდენტიფიცირებასა და შეფასებაზე, ხოლო მეორეს მხრივ მაღალი სიხშირის მონაცემების ხარისხზე (იხ. დანართი 8.1). როგორც მონაცემების ანალიზისას გამოჩნდა, მოდელებში გამოყენებული მონაცემები მყისიერად ასახავდნენ მიმდინარე პერიოდის მდგომარეობას, მაშინ როცა 2020 წლის მარტში მშპ- მკვეთრად შემცირდა, ასევე მკვეთრად შემცირდა უმეტესი მაღალი სიხშირის მონაცემიც, და აღდგენის პერიოდშიც 2021 წლის მარტში ნაწილობრივ დაბალი საბაზო ეფექტის გამო და ეკონომიკური აქტივობის გაჯანსაღებასთან ერთად უმეტესი შერჩეული მონაცემები ასევე მკვეთრ წლიურ ზრდას აფიქსირებდნენ.

თუმცა, უნდა აღინიშნოს, რომ მოცემული ანალიზისას გამოყენებული იქნა ერთი და იგივე მონაცემთა ბაზა, რაც ფილტრაციის შედეგად (იხ. მონაცემების აღწერის ქვეთავი) იქნა მიღებული. მოცემული მიდგომა შესაძლოა არ იყოს თანაბრად მომგებიანი ყველა მოდელისთვის, პირველ რიგში კი არაწრფივი მოდელებისათვის, რადგან ცვლადების შერჩევა მოხდა ძირითადად კორელაციის კოეფიციენტზე დაყრდნობით, რაც თავისთავად გულისხმობს წრფივ დამოკიდებულებას. ასევე აღსანიშნავია რომ ფილტრაციის შედეგად მკვეთრად შემცირდა ცვლადების რაოდენობა, ერთის მხრივ ამ მიდგომით DFM მოდელის ხარისხი გაიზარდა, რადგან ფაქტორებში ნაკლებად შევიდა ისეთი ცვლადები, რომელიც მაღალი მერყეობით გამოირჩეოდა, მაგრამ დაბალი კავშირი ჰქონდა მშპ-სთან, მაგრამ მეორეს მხრივ შესაძლებელია მონაცემების „აგრესიული ინჟინერიით“ ახალი ცვლადების დამატება ისე, რომ მოხდეს DFM-ის შედეგების გაუმჯობესება, მაღალი კორელაციის ცვლადების (რომლებსაც არ ექნებათ რაიმე ეკონომიკური ინტერპრეტაცია) დამატებით, მაგრამ ეს საკითხი სცდება აღნიშნული ნაშრომის მიზნებს და ცალკე კვლევის საგანს წარმოადგენს. რადგან აღნიშნული ნაშრომის მთავარ მიზანს

წარმოადგენს მოდელირების მეთოდების შედარება სხვა თანაბარ პირობებში, ამიტომ თითოეული მოდელის შედეგების გაუმჯობესება შესაძლებელია მონაცემების თითოეულ მოდელზე ინდივიდუალურად მორგებით.

6. დასკვნა

აღნიშნული კვლევა მიზნად ისახავდა ეკონომიკური ზრდის ნაუქასთისთვის თანამედროვე მონაცემებისა და მეთოდების გამოყენების შესაძლებლობების და მიზანშეწონილობის შესწავლას საქართველოს მაგალითზე.

მიღებული შედეგები ცხადყოფს, რომ შეფასების მეთოდებისა და დამატებითი მონაცემების წყაროების მოძიების შემდგომ შესაძლებელია, ყოველთვისურად „მიმდინარე“ რეჟიმში ეკონომიკურ აქტივობაზე დაკვირვება. ანალიზისას გამოჩნდა რომ მანქანური სწავლების მეთოდები, განსაკუთრებით კი რეგულარიზებული წრფივი მოდელები, კერნელის მოდელები და ნეირონული ქსელები მაღალი სიზუსტით გამოირჩევიან, ვიდრე ტრადიციული და ბენჩმარკ მიდგომები. აღნიშნული მოდელები მაღალი სიზუსტით აფასებენ, როგორც ეკონომიკური აქტივობის მასშტაბს ასევე მიმართულებას. თითქმის ყველა მოდელის შემთხვევაში შესაძლებელი იყო 2020 წლის მყისიერი ეკონომიკური ვარდნის შეფასება, რაც ასევე მიუთითებს სწრაფი მონაცემების ხარისხზე, როგორც ტრადიციული მონაცემების წყაროებიდან მიღებულ მონაცემების ასევე Google trend-ის ძიების ინდექსების შემთხვევაში. ანალიზის პროცესში გამოჩნდა, რომ შესაძლებელია Google Trend-ის მეშვეობის ისეთი მაღალი სიხშირის ცვლადების მოძიება, რომელიც კარგად და მყისიერად ასახავენ მიმდინარე ეკონომიკურ ვითარებას. მაგრამ ასევე საგულისხმოა ის ფაქტი, რომ მრავალი ასეთი მონაცემიდან ბევრი მოდელირებისას გამოუსადეგარია.

7. გამოყენებული ლიტერატურა

- Antolin-Diaz, J., Drechsel, T., & Petrella, I. (2021). Advances in Nowcasting Economic Activity: Secular Trends, Large Shocks and New Data. In *CEPR Discussion Papers* (No. 15926; CEPR Discussion Papers). C.E.P.R. Discussion Papers. <https://ideas.repec.org/p/cpr/ceprdp/15926.html>
- Austin, P., Marini, M., Sanchez, A., Simpson-Bell, C., & Tebrake, J. (2021). *Using the Google Places API and Google Trends data to develop high frequency indicators of economic activity*. Washington, DC : International Monetary Fund.
- Baffigi, A., Golinelli, R., & Parigi, G. (2004). Bridge models to forecast the euro area GDP. *International Journal of Forecasting*, 20(3), 447–460.
- Bańbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Chapter 4—Now-Casting and the Real-Time Data Flow. In G. Elliott & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 2, pp. 195–237). Elsevier. <https://doi.org/10.1016/B978-0-444-53683-9.00004-9>
- Barhoumi, K., Choi, S. M., Iyer, T., Li, J., Ouattara, F., Tiffin, A., & Yao, J. (2022). *Overcoming Data Sparsity: A Machine Learning Approach to Track the Real-Time Impact of COVID-19 in Sub-Saharan Africa* (SSRN Scholarly Paper No. 4117838). <https://papers.ssrn.com/abstract=4117838>
- Basuchoudhary, A., Bang, J., & Shughart II, W. (2020). *Predicting State Failure: Different Paths into the Abyss*. <https://doi.org/10.2139/ssrn.1731660>
- Baumeister, C., & Guérin, P. (2020). A comparison of monthly global indicators for forecasting growth. *CAMA Working Papers*, Article 2020–93. <https://ideas.repec.org/p/een/camaaa/2020-93.html>
- Bolhuis, M., & Rayner, B. (2020). *Deus Ex Machina? A Framework for Macro Forecasting with Machine Learning* (SSRN Scholarly Paper No. 3579665). <https://papers.ssrn.com/abstract=3579665>

- Bräuning, F., & Koopman, S. J. (2014). Forecasting macroeconomic variables using collapsed dynamic factor analysis. *International Journal of Forecasting*, 30(3), 572–584.
- Camacho, M., Perez-Quiros, G., & Poncela, P. (2013). *Short-term forecasting for empirical economists. A survey of the recently proposed algorithms* (Working Paper No. 1318). Banco de España. <https://econpapers.repec.org/paper/bdewpaper/1318.htm>
- Carrière-Swallow, Y., & Labbé, F. (2010). Nowcasting With Google Trends in an Emerging Market. *Working Papers Central Bank of Chile*, Article 588. <https://ideas.repec.org/p/chb/bcchwp/588.html>
- Chakraborty, C., & Joseph, A. (2017). *Machine Learning at Central Banks* (SSRN Scholarly Paper No. 3031796). <https://doi.org/10.2139/ssrn.3031796>
- Chapman, J. T. E., & Desai, A. (2021). *Macroeconomic Predictions using Payments Data and Machine Learning* (SSRN Scholarly Paper No. 3907281). <https://doi.org/10.2139/ssrn.3907281>
- Clements, M. P., & Galvao, A. B. (2009). Forecasting US output growth using leading indicators: An appraisal using MIDAS models. *Journal of Applied Econometrics*, 24(7), 1187–1206.
- Dauphin, J.-F., Dybczak, K., Maneely, M., Taheri Sanjani, M., Suphaphiphat, N., Wang, Y., & Zhang, H. (2022). *Nowcasting GDP - a Scalable Approach Using DFM, Machine Learning and Novel Data, Applied to European Economies* (SSRN Scholarly Paper No. 4082978). <https://papers.ssrn.com/abstract=4082978>
- Doz, C., Giannone, D., & Reichlin, L. (2006). *A Two-step estimator for large approximate dynamic factor models based on Kalman filtering* (THEMA Working Paper No. 2006–23). THEMA (Théorie Economique, Modélisation et Applications), Université de Cergy-Pontoise. <https://econpapers.repec.org/paper/emaworppap/2006-23.htm>
- Giannone, D., Reichlin, L., & Simonelli, S. (2009). *Nowcasting Euro Area Economic Activity in Real-Time: The Role of Confidence Indicators* [CSEF Working Paper]. Centre for Studies in Economics and

Finance (CSEF), University of Naples, Italy.

<https://econpapers.repec.org/paper/sefcsefwp/240.htm>

- Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4), 665–676.
- Ginker, T., & Suhoy, T. (2021). Nowcasting and monitoring Israeli real economic activity. *IFC Bulletins Chapters*, 55. <https://ideas.repec.org/h/bis/bisifc/55-03.html>
- Kohlscheen, E. (2021). *What Does Machine Learning Say About the Drivers of Inflation?* (SSRN Scholarly Paper No. 3949352). <https://doi.org/10.2139/ssrn.3949352>
- Kuzin, V., Marcellino, M., & Schumacher, C. (2011). MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area. *International Journal of Forecasting*, 27(2), 529–542.
- Lewis, D. J., Mertens, K., Stock, J. H., & Trivedi, M. (2022). Measuring real activity using a weekly economic index. *Journal of Applied Econometrics*, 37(4), 667–687. <https://doi.org/10.1002/jae.2873>
- Mariam Dundua & Otar Gorgodze. (2022, November). *Application of Artificial Intelligence for Monetary Policy-Making*. National Bank of Georgia. <https://nbg.gov.ge/publications/researches?page=1>
- Matheson, T. (2011). *New Indicators for Tracking Growth in Real Time* (SSRN Scholarly Paper No. 1770369). <https://papers.ssrn.com/abstract=1770369>
- Muchisha, N., Tamara, N., Andriansyah, A., & Soleh, A. (2021). Nowcasting Indonesia's GDP Growth Using Machine Learning Algorithms. *Indonesian Journal of Statistics and Its Applications*, 5, 355–368. <https://doi.org/10.29244/ijsa.v5i2p355-368>
- Narita, M. F., & Yin, R. (2018). In Search of Information: Use of Google Trends' Data to Narrow Information Gaps for Low-income Developing Countries. *IMF Working Papers*, Article 2018/286. <https://ideas.repec.org/p/imf/imfwpa/2018-286.html>

- Quah, D., & Sargent, T. J. (1993). A Dynamic Index Model for Large Cross Sections. *NBER Chapters*, 285–310.
- Richardson, A., van Florenstein Mulder, T., & Vehbi, T. (2021). Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, 37(2), 941–948.
- Robin, F. (2018). Use of Google Trends Data in Banque de France Monthly Retail Trade Surveys. *Economie et Statistique / Economics and Statistics*, 505–506, 35–63.
- Sampi Bravo, J. R. E., & Jooste, C. (2020). *Nowcasting Economic Activity in Times of COVID-19: An Approximation from the Google Community Mobility Report* (SSRN Scholarly Paper No. 3601423). <https://papers.ssrn.com/abstract=3601423>
- Sargent, T. (1989). Two Models of Measurements and the Investment Accelerator. *Journal of Political Economy*, 97(2), 251–287.
- Sargent, T., & Sims, C. (1977). *Business cycle modeling without pretending to have too much a priori economic theory* (Working Paper No. 55). Federal Reserve Bank of Minneapolis. <https://econpapers.repec.org/paper/fipfedmwp/55.htm>
- Stock, J., & Watson, M. (1989). *New Indexes of Coincident and Leading Economic Indicators* (pp. 351–409) [NBER Chapters]. National Bureau of Economic Research, Inc. <https://econpapers.repec.org/bookchap/nbrnberch/10968.htm>
- Ter-Martirosyan, A., Patnam, M., & Jung, J.-K. (2018). An Algorithmic Crystal Ball: Forecasts-based on Machine Learning. *IMF Working Papers*, Article 2018/230. <https://ideas.repec.org/p/imf/imfwpa/2018-230.html>
- Tiffin, A. (2016). *Seeing in the Dark: A Machine-Learning Approach to Nowcasting in Lebanon* (SSRN Scholarly Paper No. 2770291). <https://papers.ssrn.com/abstract=2770291>
- Watson, M. W., & Engle, R. F. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, 23(3), 385–400. [https://doi.org/10.1016/0304-4076\(83\)90066-0](https://doi.org/10.1016/0304-4076(83)90066-0)

- Woloszko, N. (2020). Tracking activity in real time with Google Trends. *OECD Economics Department Working Papers*, Article 1634. <https://ideas.repec.org/p/oec/ecoaaa/1634-en.html>
- Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)

8. დანართი

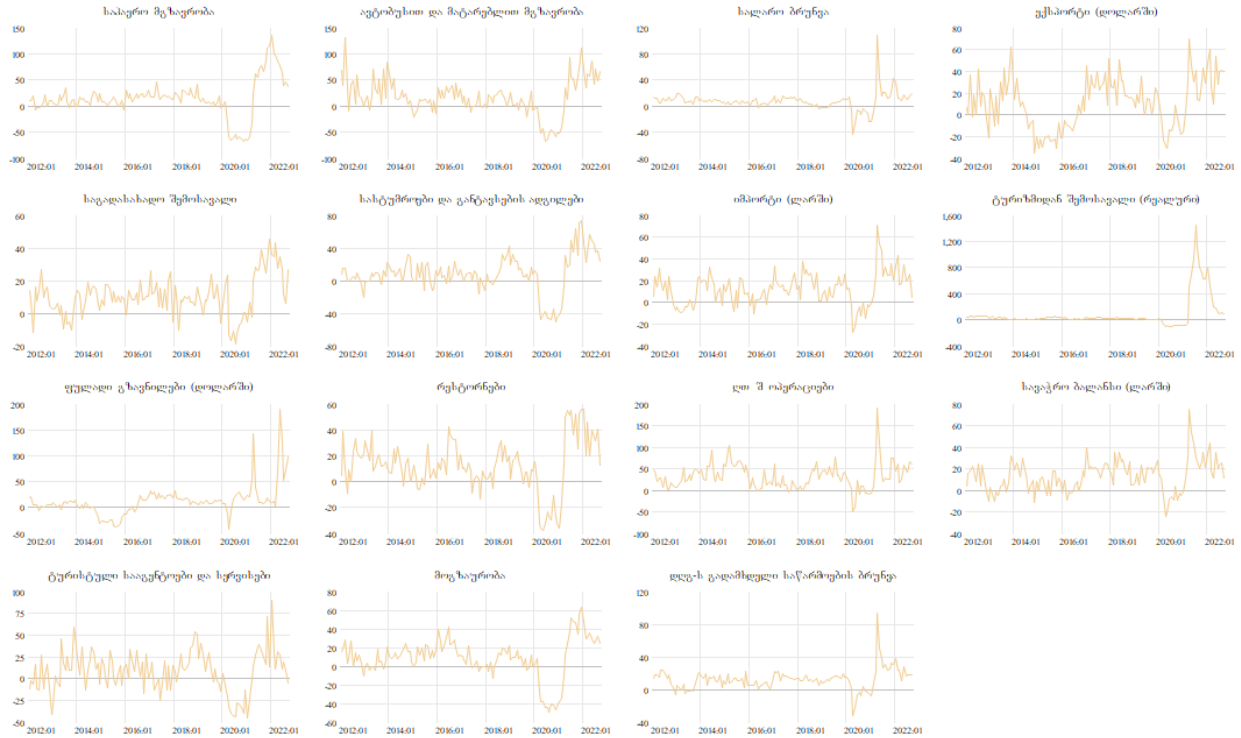
8.1 ნაშრომში გამოყენებული მონაცემები

ცხრილი 3 . ნაშრომში გამოყენებული მონაცემები

წყარო	სახელი	მონაცემის ტიპი	გამოქვეყნების თარიღი	კორელაცია მშპ-სთან
MOF	საკასო ბრუნვა	I	შემდგომი თვის პირველი რიცხვი	0.68
	საგადასახადო შემოსავლები	I	შემდგომი თვის პირველი რიცხვი	0.55
	მთავრობის ხარჯები	I	შემდგომი თვის პირველი რიცხვი	0.06
	საქონლით ვაჭრობის მონაცემები (იმპორტი)	I	შემდგომი თვის პირველი რიცხვი	0.63
	საქონლით ვაჭრობის მონაცემები (ექსპორტი)	I	შემდგომი თვის პირველი რიცხვი	0.43
	საქონლით ვაჭრობის მონაცემები (ბრუნვა)	I	შემდგომი თვის პირველი რიცხვი	0.65
	დღგ-ს ბრუნვა	II	შემდგომი თვის 15 რიცხვი	0.93
NBG	სავალუტო კურსები (USD/GEL)	I	შემდგომი თვის პირველი რიცხვი	-0.29
	სავალუტო კურსები (NEER)	I	შემდგომი თვის პირველი რიცხვი	0.27
	TIBR	I	შემდგომი თვის 2 რიცხვი	-0.09
	M2 (წინასწარი)	I	შემდგომი თვის 3 რიცხვი	0.22
	RTGS ოპერაციები	I	შემდგომი თვის 5 რიცხვი	0.62
	ტურიზმიდან შემოსავლები	II	შემდგომი თვის 15 რიცხვი	0.56
	ფულადი გზავნილები	II	შემდგომი თვის 15 რიცხვი	0.41
GeoStat	ინფლაცია (CPI)	I	შემდგომი თვის 3 რიცხვი	0.28
	საქონლით ვაჭრობის მონაცემები (იმპორტი)	II	შემდგომი თვის 15 რიცხვი	0.62
	საქონლით ვაჭრობის მონაცემები (ექსპორტი)	II	შემდგომი თვის 15 რიცხვი	0.75
	საქონლით ვაჭრობის მონაცემები (ბრუნვა)	II	შემდგომი თვის 15 რიცხვი	0.76
Google	Travel index	I	შემდგომი თვის 3 რიცხვი	0.58
	Air travel	I	შემდგომი თვის 3 რიცხვი	0.63
	Bus and rail	I	შემდგომი თვის 3 რიცხვი	0.57
	Car rental and taxi service	I	შემდგომი თვის 3 რიცხვი	0.25
	Cruises and charters	I	შემდგომი თვის 3 რიცხვი	0.27
	Hotels and Accommodations	I	შემდგომი თვის 3 რიცხვი	0.60
	Tourist destinations	I	შემდგომი თვის 3 რიცხვი	0.32
	Travel agencies and services	I	შემდგომი თვის 3 რიცხვი	0.50
	Travel guides and travelogues	I	შემდგომი თვის 3 რიცხვი	0.34
	Auto index	I	შემდგომი თვის 3 რიცხვი	0.23
	Vehicle licensing and registration	I	შემდგომი თვის 3 რიცხვი	0.21
	Vehicle parts and accessories	I	შემდგომი თვის 3 რიცხვი	0.35
	Advertising and marketing	I	შემდგომი თვის 3 რიცხვი	-0.23
	Business news	I	შემდგომი თვის 3 რიცხვი	-0.21
	Construction and maintenance	I	შემდგომი თვის 3 რიცხვი	0.19
	Energy and utilities	I	შემდგომი თვის 3 რიცხვი	0.28
	Metals and mining	I	შემდგომი თვის 3 რიცხვი	0.14
	small business	I	შემდგომი თვის 3 რიცხვი	-0.38
	transportation and logistics	I	შემდგომი თვის 3 რიცხვი	0.21
	Financial planing	I	შემდგომი თვის 3 რიცხვი	0.27
	Investing	I	შემდგომი თვის 3 რიცხვი	-0.18
	non-alcoholic beverages	I	შემდგომი თვის 3 რიცხვი	0.28
	Restaurants	I	შემდგომი თვის 3 რიცხვი	0.67
	Apparel	I	შემდგომი თვის 3 რიცხვი	0.28
	Consumer resources	I	შემდგომი თვის 3 რიცხვი	0.22
	Weddings	I	შემდგომი თვის 3 რიცხვი	0.24
	luxury goods	I	შემდგომი თვის 3 რიცხვი	0.17
tobacco products	I	შემდგომი თვის 3 რიცხვი	0.33	
clubs and nightlife	I	შემდგომი თვის 3 რიცხვი	0.34	
Events and listings	I	შემდგომი თვის 3 რიცხვი	0.33	
concerts and music festivals	I	შემდგომი თვის 3 რიცხვი	0.32	

წყარო : ფინანსთა სამინისტრო, შემოსავლების სამსახური, ეროვნული ბანკი, საქსტატი, Google trends

გრაფიკი 6. მოდელში გამოყენებული მონაცემები.



წყარო : ფინანსთა სამინისტრო, შემოსავლების სამსახური, საქსტატი, ეროვნული ბანკი, Google Trends

8.2 მანქანური სწავლების მოდელები

8.2.a რეგულარიზებული წრფივი რეგრესიები (LASSO, Ridge, Elastic Net)

რეგულარიზების მეთოდის საშუალებით ხდება ცვლადების კოეფიციენტების შემცირება (shrinkage) ნულისკენ, რითაც ხდება მოდელის overfitt-ის თავიდან აცილება. რეგულარიზაცია შესაძლებელია გამოყენებული იყოს უმეტეს მანქანური სწავლებით მოდელირებისას, მაგრამ ის ყველაზე ხშირად გამოყენებადი წრფივი მოდელირებისას, სადაც მისი გამოყენებით ბეტა კოეფიციენტების ნულისკენ შემცირება ხდება. ყველაზე პოპულარული წრფივი რეგულარიზაციის მოდელებია Ridge, least absolute shrinkage and selector operator (LASSO) და Elastic Net რეგრესიები.

რეგულარიზებული რეგრესია, ისევე როგორც უმცირეს კვადრატთა მეთოდი ახდენს დანაკარგის ფუნქციის მინიმიზაციას, რომელიც მცირედით არის მოდიფიცირებული. თუ წრფივი მოდელს აქვს შემდეგი სახე:

$$y_t = x_t' \beta + \varepsilon_t$$

სადაც y_t წარმოადგენს დამოკიდებულ ცვლადს, ხოლო x_t წარმოადგენს $n \times 1$ ამხსნელი ცვლადების ვექტორს, ხოლო ε_t ნარჩენობითი წევრია. ამ განტოლების პარამეტრების შეფასებისთვის ტრადიციული უმცირეს კვადრატთა მეთოდი ახდენს შემდეგი დანაკარგის ფუნქციის მინიმიზებას

$$\hat{\beta} = \min [\sum_1^t (y_t - x_t' \beta)^2]$$

რაც შინაარსობრივად წამოადგენს ნარჩენობითი წევრების კვადრატთა ჯამის შემცირებას. რეგულარიზების მეთოდი კი, იმისათვის რომ არ მოხდეს overfitt, დანაკარგის ფუნქციის მოდიფიკაციას ახენს ფუნქციაში „ჯარიმის“ (penalty) შეტანით.

LASSO მეთოდის მიხედვით უმცირეს კვადრატთა მეთოდის დანაკარგის ფუნქციას ემატება კოეფიციენტის აბსოლუტური მაჩვენებელი როგორც „ჯარიმა“, რასაც ასევე L1 რეგულარიზაცია ეწოდება. მოდიფიცირებულ ფუნქციას აქვს შემდეგი სახე:

$$\hat{\beta} = \min \left[\sum_1^t (y_t - x_t' \beta)^2 + \lambda_1 |\beta| \right]$$

Ridge მეთოდის მიხედვით უმცირეს კვადრატთა მეთოდის დანაკარგის ფუნქციას ემატება კოეფიციენტის აბსოლუტური მაჩვენებლის კვადრატი როგორც „ჯარიმა“, რასაც ასევე L2 რეგულარიზაცია ეწოდება. მოდიფიცირებულ ფუნქციას აქვს შემდეგი სახე:

$$\hat{\beta} = \min [\sum_1^t (y_t - x_t' \beta)^2 + \lambda_2 |\beta|^2]$$

Elastic Net კი წარმოადგენს L1 და L2 რეგულარიზაციების გაერთიანებას, და დანაკარგის ფუნქცია იღებს შემდეგ სახეს

$$\hat{\beta} = \min [\sum_1^t (y_t - x_t' \beta)^2 + \lambda_1 |\beta| + \lambda_2 |\beta|^2]$$

λ_1 და λ_2 წარმოადგენს ჰიპერპარამეტრებს, რომელთა მოდელისთვის მორგებაც (Tuning) ხდება სხვადასხვა რაოდენობრივი მეთოდით, როგორცაა Grid და Random ძიების მეთოდები. ნაშრომში ჰიპერპარამეტრების მორგებისთვის გამოყენებულია კროს-ვალიდაცია, ის პარამეტრები

რომლებიც კროს-ვალიდაციისას, ვალიდაციის სეტზე (Validation set) უკეთესი შედეგით გამოირჩევა არჩეულია საბოლოო მოდელისთვის.

8.2.b კერნელის მეთოდები (Support Vector Machine (SVM))

კერნელის მეთოდები რომელიც მოიცავს Support Vector Regression (SVR), ახდენს კომპლექსური, არაწრფივი ურთიერთკავშირების მოდელირებას, თავდაპირველად მონაცემების მაღალ განზომილებაში (higher dimension) გადაყვანით, ხოლო შემდეგ წრფივი მორგებული სიბრტყის (hyperplane) მეშვეობით ხდება კოეფიციენტების შეფასება. არსებობს რამდენიმე გზა თუ როგორ შეიძლება შეფასდეს აღნიშნული მორგებული სიბრტყეები, მაგრამ SVR-ის შემთხვევაში მისი პოვნისას გამოყენებულია არა ნარჩენობითი წევრების კვადრატთა ჯამი (უმცირეს კვადრატთა მეთოდის დანაკარგის ფუნქცია), არამედ ნარჩენობითი წევრების აბსოლუტური მნიშვნელობების ჯამი, მაგრამ მხოლოდ იმ შემთხვევაში თუ ის წინასწარ განსაზღვრული ზღვარს აღემატება. რაც ნიშნავს, რომ ბეტა კოეფიციენტების შეფასებისას ხდება მათი სიდიდის შემცირება, ისე რომ ნარჩენობითი წევრები გარკვეულ დიაპაზონში იყვნენ მოქცეულნი.

წრფივი რეგრესიის მსგავსად დამოკიდებული ცვლადი მოდელირებულია შემდეგნაირად:

$$f(x_t) = x_t' \beta + \varepsilon_t$$

SVM ალგორითმის მიზანია იპოვოს ფუნქცია $f(x_t)$, ისე რომ ის რაც შეიძლება ბრტყელი იყოს და ამავდროულად ნარჩენობითი წევრები იყვნენ წინასწარ განსაზღვრულ დიაპაზონში. SVM ახდენს დანაკარგის ფუნქციის მინიმიზებას ბეტა და დაკვირვება სპეციფიკური მუდმივების ζ_t და ζ_t^* კომბინაციით.

$$\frac{1}{2} \beta' \beta + C \sum_{t=1}^T (\zeta_t + \zeta_t^*)$$

დაქვემდებარებული შემდეგ შეზღუდვებს, $y_t - f(x_t) \leq \varepsilon + \zeta_t$

$$\text{და } f(x_t) - y_t \leq \varepsilon + \zeta_t^*$$

შესაბამისად, ყოველ დაკვირვების წერტილზე, ნარჩენობითი წევრი არის შეზღუდული $[-\varepsilon, \varepsilon]$, და შესაძლებელია ამ დიაპაზონის დარღვევა ζ_t და ζ_t^* სიდიდით. მუდმივა C წარმოადგენს ჰიპერპარამეტრს, რომელიც კონტროლს უწევს თუ რამდენად მოხდეს ნარჩენობითი წევრების

მნიშვნელობის შემცირება და overfitt, თუ რომელ მხარეს უფრო გადაიხაროს. თუ $C = 0$, მაშინ მოხდება ყველაზე მარტივი ფორმის მორგებული სიბრტყის აგება, სადაც ყველა დაკვირვების წერტილი მისაღებ დიაპაზონში იმყოფება. C -ს ზრდასთან ერთად იზრდება მორგებული წრფის კომპლექსურობაც, და ნარჩენობითი წევრების მნიშვნელობა ნულთან ახლოვდება, რაც საფრთხეს აჩენს overfitt-ინგის. SVM ანალიზისას შესაძლებელია კერნელის ფუნქციის შეცვლა და ის შეიძლება იყოს - წრფივი, სხვადასხვა ხარისხის პოლინომიალი, სიგმოიდ ფუნქცია და ა.შ. ყველაზე პოპულარულ მეთოდს კი წარმოადგენს გაუსის რადიალური ბაზის ფუნქციის (Gaussian Radial Basis Function (RBF)) კერნელი

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

8.2.c Ensemble methods (Decision trees, Random Forests, Gradient boosting, Adaptive boosting)

გადაწყვეტილების ხეები (Decision trees) წარმოადგენენ არაპარამეტრულ მოდელებს, სადაც კოეფიციენტების შეფასების მაგივრად ხდება შერჩევის ორ ნაწილად გაყოფა. შესაბამისად გადაწყვეტილების ხეები შეგვიძლია განვიხილოთ, როგორც დიაგრამის მსგავსი სტრუქტურები, რომლებიც შექმნილია კონკრეტული გამოსავლის პროგნოზირებისთვის. მაგალითად ავიღოთ დიაგრამა, სადაც თითოეული დონე არის კითხვა ორობითი (binary) დიახ ან არა პასუხით (მაგალითად, „დღეს ბრუნვის ზრდა 20%-ზე მაღალია?“) რასაც მოჰყვება სხვა დონეები ორობითი პასუხებით. დიაგრამის მიდევნება და კითხვებზე სათითაოდ პასუხის გაცემა, საბოლოოდ იძლევა საწყისი პრობლემის გადაჭრას. რეგრესიული ანალიზისთვის პასუხის მიღება ხდება შემდეგნაირად, ვთქვათ გვაქვს შერჩევის მხოლოდ ორ ნაწილად გაყოფა დღეს ბრუნვების მიხედვით, რის შედეგადაც მიიღება ფოთოლი, პირველი ფოთოლი აკმაყოფილებს კითხვას დღეს ბრუნვის 20%-ზე მეტობას, რის ფოთლის მნიშვნელობა განისაზღვრება როგორც მშპ-ს ზრდის (დამოკიდებული ცვლადი რის პროგნოზირებასაც ვცდილობთ) საშუალო არითმეტიკული, მაშინ როცა დღეს ზრდა 20%-ს აღემატებოდა. მეორე ფოთლის მნიშვნელობაც ანალოგიურად ხორციელდება, ოღონდ იმ განსხვავებით რო ხდება მშპ-ს საშუალოს აღება იმ შემთხვევაში თუ დღეს ბრუნვა 20%-ზე ნაკლები იყო. შერჩევის დაყოფის ზრდასთან ერთად იზრდება მოდელის კომპლექსურობაც და რისკიც იმისა რომ მოხდეს overfitt, რაც მოდელს საშუალებას არ მისცემს მოახდინოს ცვლადებს შორის ურთიერთობის განვრცობა მომავალში. მაგალითად თუ ავიღებთ მოდელს ერთი ამხსნელი ცვლადით (დღეს ბრუნვა) და მოვახდენთ შერჩევის იმდენჯერ დაყოფას რამდენი დაკვირვების წერტილიც დღეს ბრუნვას გააჩნია, მაშინ მივიღებთ დღეს ყოველი ბრუნვის მაჩვენებლისთვის მშპ-ს ერთ მნიშვნელობას, რაც რატემაუნდა მცდარია და

მოდელი ახალ მონაცემებზე ვერ იმუშავებს, რაც overfitt-ინგის კლასიკური მაგალითია. რეგრესული გადაწყვეტილების ხის ფორმალური გამოსახვა შემდეგნაირია:

$$\hat{f}(x) = \sum_{m=1}^M \hat{c}_m I(x \in R_m)$$

სადაც $I(\cdot)$ არის ინდიკატორის ფუნქცია და $\hat{c}_m = \text{avg}(y_i | x \in R_m)$, და R_m წარმოადგენს იმ ჯგუფებს (ფოთლებს), რაც დაყოფის შედეგად მიიღება. ალგორითმის მიზანია იპოვოს ისეთი ოპტიმალური R_m და \hat{c}_m , რომლის დროსაც მოხდება ნარჩენობითი წევრების კვადრატების ჯამის მინიმიზება.

შემთხვევითი ტყე (Random Forest – (RF)) წარმოადგენს გადაწყვეტილების ხეებისგან მიღებული პროგნოზების კომბინაციებს. ვინაიდან მხოლოდ ერთი ხისგან შემდგარი მოდელი მიდრეკილია overfitt-ისკენ და ასევე ის მეტწილად დამოკიდებულია ლოკალურ ოპტიმიზაციაზე და არა გლობალურზე, შემთხვევითი ტყე სწორედ ამ პრობლემებზე კონცენტრირდება. RF იყენებს bootstrap¹⁰ -ის აგრეგაციას (bagging¹¹), რის შემდეგაც იქმნება გადაწყვეტილებების ხეების „ტყე“, თითოეული გადაწყვეტილების ხე შედგენილია შემთხვევითად შერჩეული მონაცემებიდან, რომელიც bagging-ის შემდგომ მიიღება. რეგრესიული ანალიზისთვის საბოლოო შედეგი წარმოადგენს ყველა გადაწყვეტილების ხის შედეგის საშუალო არითმეტიკულს.

Gradient boosting RF-ის მსგავსად იყენებს სხვადასხვა გადაწყვეტილების ხეების შედეგებს, მაგრამ მთავარი განსხვავება ისაა, რომ ერთდროულად ყველა მოდელის საშუალოს აღების მაგივრად, Gradient boosting მეთოდი აგებს გადაწყვეტილებების ხეებს ერთმანეთის მიყოლებით, სადაც ყოველი შემდგომი ხე იყენებს წინა ხის პროგნოზებს და ცდილობს შეამციროს ნარჩენობითი წევრი. საბოლოო შედეგი წარმოადგენს პროცესში მონაწილე ყველა გადაწყვეტილების ხის შედეგის საშუალოს. პრაქტიკაში გავრცელებულია Gradient boosting სამი ალგორითმი, XGboost (Extreme gradient boosting), LightGBM (light gradient boosting machine) და Catboost (Categorical boosting). ეს ორი უკანასკნელი შემუშავებული იქნა Microsoft-ისა და Yandex-ის მიერ. XGboost პრაქტიკაში რთულად გამოსაყენებელია, რადგან მისი შეფასებისთვის დიდი დროა საჭირო, მაშინ როცა LightGBM და Catboost შეფასება მნიშვნელოვნად სწრაფად ხდება.

¹⁰ Bootstrap წარმოადგენს სტატისტიკურ პროცედურას როდესაც მონაცემებიდან ხდება შერჩევის შემთხვევითობის წესით შერჩევა, რომელთა მეშვეობითაც ყალიბდება ახალი მონაცემები.

¹¹ Bagging იგივე bootstrap აგრეგაცია, რის შედეგადაც ხდება დამატებითი მონაცემების შედგენა უკვე არსებული მონაცემებიდან შემთხვევითობის პრინციპით.

Adaptive boosting იყენებს მცირედით განსხვავებულ მეთოდს, მაშინ როდესაც Gradient boosting წინა მოდელის შეცდომების „გამოსწორებას“ ცდილობს, Adaptive boosting მიდგომა კი ყოველი შემდგომი გამეორებისას იმ გადაწყვეტილებების ხეებს, რომლებსაც ცუდი შედეგი ქონდათ (weak learners) მეტ წონას ანიჭებს, რათა მოხდეს მათი გარდაქმნა „სუსტი მსწავლელებიდან“ „ძლიერ მსწავლელებად“. საბოლოო შედეგი კი წარმოადგენს შეწონილ საშუალოს სადაც წონები განაწილებულია მოდელების სიზუსტის მიხედვით.

ყველა ზემოთ აღნიშნული მოდელის მთავარ პრობლემას წარმოადგენს ორი გარემოება. პირველი ესაა overfitt-ისკენ მიდრეკილება, იმ შემთხვევაშიც კი თუ ვიყენებთ მეტად კომპლექსურ მეთოდებს (Gradient boosting, Adaptive boosting). მეორე გარემოებას კი წარმოადგენს რომ მოდელის შედეგების ექსტრაპოლაცია შეზღუდული შესაძლებლობებისაა. თუ მონაცემები რომელთა მეშვეობითაც მოხდა მოდელის შეფასება კარგად არ წარმოადგენენ პოპულაციას, მაშინ მოდელი შეზღუდულია იმ შერჩევის მნიშვნელობებით რაც შეფასებისას იქნა გამოყენებული, ვინაიდან საბოლოო შედეგი წარმოადგენს დამოკიდებული ცვლადის საშუალოებს და არა რაიმე პარამეტრისა და ცვლადის კომბინაციას. შესაბამისად მსგავსი მოდელები დიდი წარმატებით გამოირჩევიან კლასიფიკაციისთვის, სადაც აღბათობაა შესაფასებელი, ხოლო რეგრესიული ანალიზისას მათ არ შეუძლიათ outlier-ების პროგნოზირება/ნაუქასთი, თუ რაიმე სახის outlier შეფასებისას არ იღებდა მონაწილეობას. ამიტომ თუ მშპ-ს შემცირება/ზრდა ნახტომისებურია და ისტორიულ მაქსიმუმს შეადგენს, რომელიც წინა მაქსიმუმს ბევრად აღემატება, აღნიშნული მოდელებიდან მიღებული შედეგებიც დაშორებული იქნება რეალური სურათისგან. მაგრამ ამ ტიპის მოდელების გამოყენება ნორმალურ პერიოდებში, როდესაც მერყეობა ნაკლებია მიზანშეწონილია.

8.2.d ნეირონული ქსელი (Neural Network)

ნეირონული ქსელი (Neural Network (NN)) წარმოადგენს მარავალშირიან არაწრფივ¹² მეთოდს, რომელიც აკავშირებს ამხსნელ ცვლადებს ასახსნელ ცვლადებთან. თითოეული შრე შედგება კვანძებისგან, რომელსაც ასევე ხელოვნურ ნეირონებს უწოდებენ. თითოეული კვანძი იღებს

¹² ყველაზე მარტივი ფორმით როდესაც ნეირონული ქსელი შედგება მხოლოდ ორი შრისგან, რომელსაც მხოლოდ გააჩნია საწყისი მონაცემების და საბოლოო შედეგის შრეები, და არცერთი დამალული შრე არ მონაწილეობს მოდელში, ხოლო ის ერთადერთი აქტივაციის ფუნქცია რომელიც პირველ შრეს საბოლოოსთან აკავშირებს წრფივი ფუნქციაა, ასეთი ნეირონული ქსელი წარმოადგენს მარტივ წრფივ მოდელს.

ინფორმაციას თავდაპირველი მონაცემებიდან, ან წინა შრის კვანძებიდან. თითოეულ კვანძის შედეგს წარმოადგენს მიღებული მონაცემების შეწონილ ჯამს (ხშირად გამოიყენება მარტივი წრფივი მოდელი, რომელიც გარდაქმნის მიღებულ მონაცემებს ერთ შედეგად), რომელიც შემდგომ არაწრფივი ფუნქციით¹³ ტრანსფორმირდება (აქტივირდება), რათა მოხდეს ახალი შედეგის ფორმირება. აქტივაციის ფუნქციის ხშირად გამოყენებადი ფორმებია,

- გასწორებული წრფივი ერთეული (Rectified linear unit (ReLU)) - $f(z) = \max(0, z)$
- ლოჯისტიკური ფუნქცია - $f(z) = \frac{1}{(1+e^{-z})}$
- ჰიპერბოლური ტანგენსი - $f(z) = \frac{e^z - e^{-z}}{(e^z + e^{-z})}$

ნეირონული ქსელის სტრუქტურიდან გამომდინარე, შესაძლებელი არაა რომ მოპოვებული იქნას პარამეტრების შეფასების უნიკალური გზა (closed form solution). ამიტომ ნეირონული ქსელის პარამეტრების შეფასებისას გამოიყენება გრადიენტზე დაშვების სტოქასტური მეთოდი (Stochastic gradient descent), რომლის დროსაც პირველ ეტაპზე პარამეტრების მნიშვნელობებზე კეთდება დაშვება, მეორე ეტაპზე ითვლება დანაკარგის ფუნქციის მნიშვნელობა ამ შემთხვევით აღებული მნიშვნელობების მიმართ, ხოლო მესამე ეტაპზე ხდება ამ პარამეტრების მნიშვნელობების ცვლილება ისე რომ მოხდეს დანახარჯის ფუნქციის შემცირება, რასაც რეგრესიული ანალიზისას, ისევე როგორც სხვა მეთოდების ხშირ შემთხვევაში წარმოადგენს ნარჩენობითი წევრების კვადრატთა ჯამს, მაგრამ შესაძლებელია დანაკარგის ფუნქციის სხვა ფორმით განსაზღვრა, მაგრამ ეს არ ცვლის ამ ფორმის (რასაც უწოდებენ feedforward Neural Network-ს) ნეირონული ქსელის მუშაობის პრინციპს.

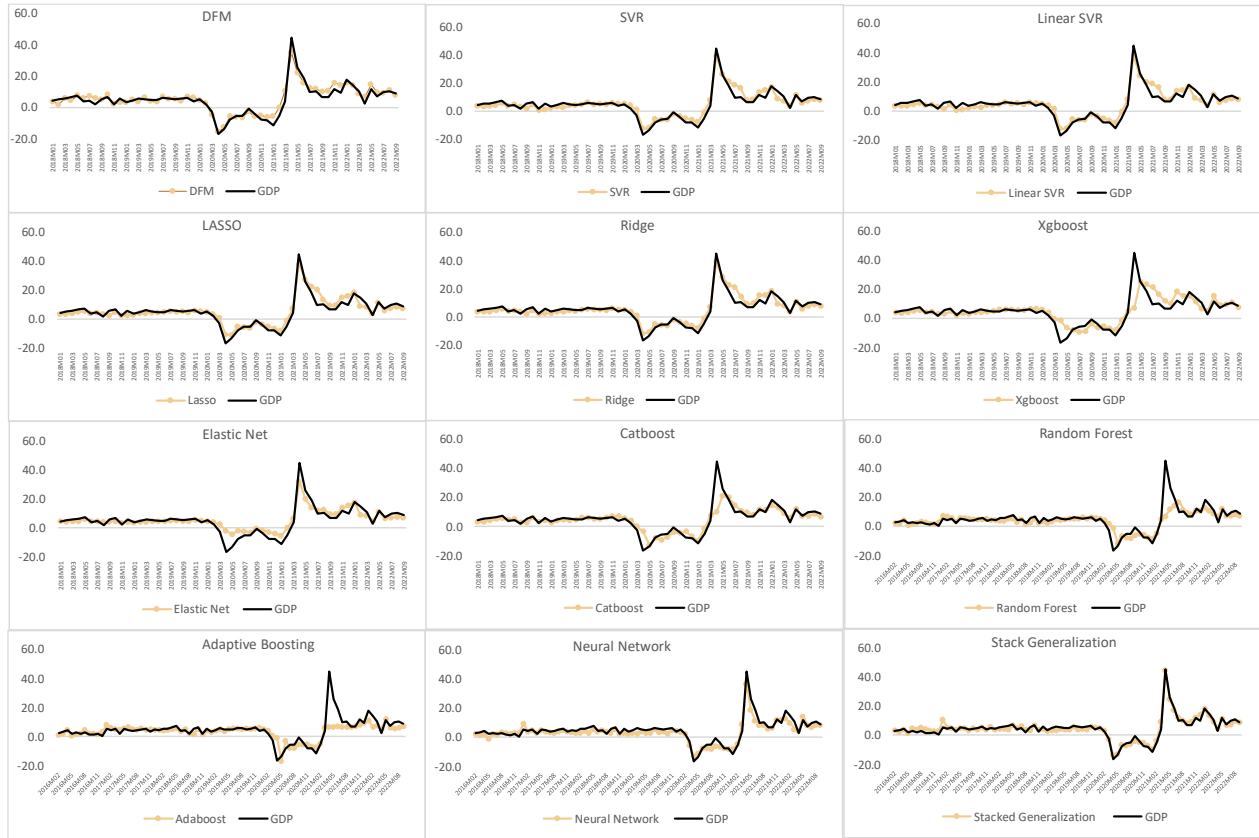
სტანდარტული ფორმით, feedforward ნეირონული ქსელი არის მთლიანად დაკავშირებული, რაც ნიშნავს რომ ყოველი შრის ყოველი კვანძი დაკავშირებულია შემდეგი შრის ყოველ კვანძთან. ხოლო საბოლოო შრე წარმოადგენს ერთ კვანძიან შრეს¹⁴ სადაც ხდება საბოლოო შედეგის მიღება.

¹³ რეგრესიული ანალიზის დროს, საჭიროა რომ საბოლოო შრე რომელიც მხოლოდ ერთი კვანძისგან შედგება (ამხსნელი ცვლადის საბოლოო მოდელით შეფასებული მნიშვნელობა) გარდაქმნილი იყოს წრფივი აქტივაციის ფუნქციით, დანარჩენ შემთხვევაში თუ ყველა შრის ტრანსფორმაციის (აქტივაციის) ფუნქცია წრფივი იქნება, მაშინ ეს ნიშნავს რომ თითოეული შრე და საბოლოოდ შედეგი წარმოადგენს პირველი შრის წრფივ კომბინაციას, რაც აზრს უკარგავს დამატებითი შრეების დამატებას, რითაც მოდელის ამხსნელი უნარი არ იზრდება.

¹⁴ ისეთი ანალიზისას, როდესაც საჭიროა ერთდროულად რამდენიმე შედეგის მიღება, შესაძლებელია ბოლო შრე შეიცავდეს იმდენ კვანძს, რამდენი ცვლადის მოდელირებაც ხდება. ყველაზე ხშირად ეს მიდგომა გამოიყენება სურათების კლასიფიკაციისას, როდესაც მიღებული შედეგი წარმოადგენს ალბათობების განაწილებას, თუ რა ალბათობით წარმოადგენს მოცემული სურათი ამა თუ იმ კატეგორიას. რის დროსაც აქტივაციის ფუნქცია წარმოადგენს სოფთმაქსი (Softmax) - $f(z_i) = \frac{e^{z_i}}{\sum_{j=1}^N e^{-z_j}}$

8.3 მოდელების შედეგები

გრაფიკი 7. მოდელის შედეგები მეორე ტიპის მონაცემებისთვის



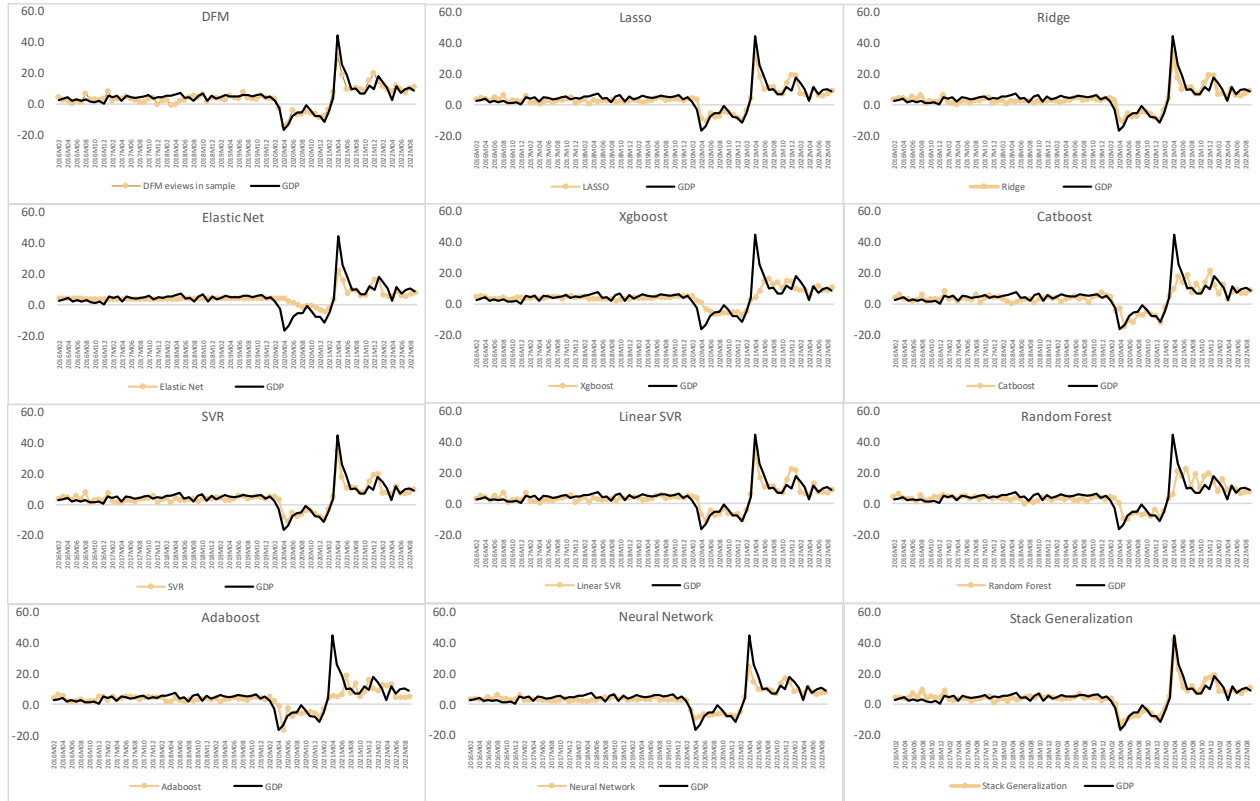
წყარო : ავტორი გაანგარიშებები

ცხრილი 4. მეორე ტიპის მონაცემების ნაუქასთინგის სიზუსტის შეფასება

მოდელები	მეორე ტიპის მონაცემები		AR(1)-თან მიმართებაში	
	RMSE	MDA	RMSE	MDA
AR(1)	5.71	66.2%	1.00	1.00
DFM views in sample	3.40	71.6%	0.60	0.92
LASSO	2.58	76.5%	0.45	0.87
Ridge	2.55	76.5%	0.45	0.87
Elastic Net	3.33	67.6%	0.58	0.98
Xgboost	5.28	70.6%	0.93	0.94
Catboost	4.56	73.5%	0.80	0.90
SVR	2.48	73.5%	0.43	0.90
Linear SVR	2.51	70.6%	0.44	0.94
Random Forest	5.23	70.6%	0.92	0.94
Adaboost	5.68	64.7%	1.00	1.02
Neural Network	2.58	66.2%	0.45	1.00
Stack Generalization	3.34	75.0%	0.59	0.88
მოდელების საშუალო	3.00	57.4%	0.53	1.15

წყარო: ავტორის გაანგარიშებები

გრაფიკი 8. მოდელის შედეგები პირველი ტიპის მონაცემებისთვის



წყარო: ავტორის განგარიშებები

ცხრილი 5. პირველი ტიპის მონაცემების ნაუქასთინგის სიზუსტის შეფასება

მოდელები	პირველი ტიპის მონაცემები		AR(1)-თან მიმართებაში	
	RMSE	MDA	RMSE	MDA
AR(1)	5.71	66.2%	1.00	1.00
DFM eviws in sample	4.45	70.1%	0.78	0.94
LASSO	3.26	70.6%	0.57	0.94
Ridge	3.25	69.1%	0.57	0.96
Elastic Net	5.04	63.2%	0.88	1.05
Xgboost	6.04	64.7%	1.06	1.02
Catboost	5.38	66.2%	0.94	1.00
SVR	3.21	69.1%	0.56	0.96
Linear SVR	3.42	69.1%	0.60	0.96
Random Forest	5.94	63.2%	1.04	1.05
Adaboost	6.31	63.2%	1.10	1.05
Neural Network	3.74	61.8%	0.65	1.07
Stack Generalization	4.98	69.1%	0.87	0.96
მოდელების საშუალო	4.00	69.1%	0.70	0.96

წყარო: ავტორის განგარიშებები