Kauno technologijos universitetas

Informatikos fakultetas

Kelionių planavimo internetinės svetainės efektyvumo didinimas grįstas vartotojų elgsenos analize ir prognozavimu

Baigiamasis magistro studijų projektas

Gytis Baravykas

Projekto autorius / autorė

dr. Agnė Paulauskaitė-Taraškevičienė

Vadovas / Vadovė

Kaunas, 2024

Kauno technologijos universitetas

Informatikos fakultetas

Kelionių planavimo internetinės svetainės efektyvumo didinimas grįstas vartotojų elgsenos analize ir prognozavimu

Baigiamasis magistro studijų projektas

Dirbtinio intelekto informatika (6211BX007)

Gytis Baravykas

Projekto autorius / autorė

dr. Agnė Paulauskaitė-Taraškevičienę

Vadovas / Vadovė

dr. Martynas Patašius

Recenzentas / Recenzentė

Kaunas, 2024

Kauno technologijos universitetas

Informatikos fakultetas

Gytis Baravykas

Kelionių planavimo internetinės svetainės efektyvumo didinimas grįstas vartotojų elgsenos analize ir prognozavimu

Akademinio sąžiningumo deklaracija

Patvirtinu, kad:

1. baigiamąjį projektą parengiau savarankiškai ir sąžiningai, nepažeisdama(s) kitų asmenų autoriaus ar kitų teisių, laikydamasi(s) Lietuvos Respublikos autorių teisių ir gretutinių teisių įstatymo nuostatų, Kauno technologijos universiteto (toliau – Universitetas) intelektinės nuosavybės valdymo ir perdavimo nuostatų bei Universiteto akademinės etikos kodekse nustatytų etikos reikalavimų;

2. baigiamajame projekte visi pateikti duomenys ir tyrimų rezultatai yra teisingi ir gauti teisėtai, nei viena šio projekto dalis nėra plagijuota nuo jokių spausdintinių ar elektroninių šaltinių, visos baigiamojo projekto tekste pateiktos citatos ir nuorodos yra nurodytos literatūros sąraše;

3. įstatymų nenumatytų piniginių sumų už baigiamąjį projektą ar jo dalis niekam nesu mokėjęs (-usi);

4. suprantu, kad išaiškėjus nesąžiningumo ar kitų asmenų teisių pažeidimo faktui, man bus taikomos akademinės nuobaudos pagal Universitete galiojančią tvarką ir būsiu pašalinta(s) iš Universiteto, o baigiamasis projektas gali būti pateiktas Akademinės etikos ir procedūrų kontrolieriaus tarnybai nagrinėjant galimą akademinės etikos pažeidimą.

Gytis Baravykas

Patvirtinta elektroniniu būdu

Kauno technologijos universitetas

Informatikos fakultetas

Baigiamojo bakalauro / magistro projekto užduotis (pagal poreikį)

Projekto tema _____

Reikalavimai ir sąlygos (tikslinti pavadinimą pagal poreikį)	

Vadovas / Vadovė _____

(vadovo pareigos, vardas, pavardė, parašas) (data)

Gytis Baravykas. Kelionių planavimo internetinės svetainės efektyvumo didinimas grįstas vartotojų elgsenos analize ir prognozavimu. Bakalauro / Magistro / Profesinių studijų / Gretutinės krypties studijų (pasirinkite) baigiamasis projektas / vadovas / vadovė dr. Agnė Paulauskaitė-Tarasevičienė; Kauno technologijos universitetas, Informatikos fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Informatikos mokslai, Informatika (B01).

Reikšminiai žodžiai: laiko eilutės, prognozavimas, duomenų analizė, kelionių paieška, vartotojo elgsenos šablonai.

Kaunas, 2024. 59 p.

Santrauka

Prognozuoti vartotojų pasirinkimus yra vis aktualesnė užduotis sparčiai besiplečiančiame internetinės prekybos tinkle. Problema tampa ypač svarbi dėl augančio tiekėjų skaičiaus, plataus paslaugų pasirinkimo ir kintančių vartotojų poreikių. Šio darbo tikslas – ištirti skirtingus statistinius ir dirbtinio intelekto modelius kelionių paieškų kiekio prognozavimui. Darbe analizuojami laiko eilučių regresijos uždaviniai, ypatingą dėmesį skiriant skrydžių paklausos prognozavimo sprendimams. Analizuojami panašių tyrimų sprendimai, įskaitant ARIMA statistinį modelį, tiesinius neuroninius tinklus, ilgalaikės trumpalaikės atminties tinklus (LSTM) ir Sequence-to-Sequence (Seq2Seq) modelius. Atliekama vartotojų duomenų analizė siekiant nustatyti vartotojų elgsenos šabloninius bruožus ieškant kelionių. Išskiriamos kelionių paieškų duomenų savybės tokios kaip keletas laiko dimensijų, geografinio tipo objektai ir šių savybių užkodavimo sprendimai. Darbe nagrinėjami du sprendimai prognozuoti paieškų kiekius: (1) transformuoti duomenis pagal skrydžio išvykimo datą, ir (2) pagal skrydžio paieškos datą. Tyrimams naudojami ARIMA, Prophet, tiesiniai daugiasluoksniai neuroniniai tinklai, LSTM ir Seq2Seq autoenkoderio modeliai. Atliekami eksperimentai, kuriuose modeliai prognozuoja 1,7 arba 30 laiko žingsnių į priekį, įvairiems oro uostų deriniams. Rezultatai, gauti prognozuojant pagal išvykimo datą, parodė, kad tiksliausiai pateikia prognozes Seq2Seq modelis su 52,3% MAPE paklaida. LSTM modelio prognozavimo paklaida siekė 59,7% MAPE, o ARIMA modelio - 56% MAPE. Eksperimentuose, vykdytuose pagal paieškos datą, pastebėta, kad geografinius objektus tinka užkoduoti koordinatėmis. Eksperimentiniai rezultatai, parodė, kad vieno laiko žingsnio prognozės neuroninių tinklų modeliuose buvo gana tikslūs, su MAPE paklaida nuo 22% iki 25%. Kelių laiko žingsnių prognozavimo uždaviniuose LSTM ir Seq2Seq modeliai pasiekė geresnius rezultatus lyginant su tiesiniu daugiasluoksniu neuroniniu, vidutiniškai 10-30% MAPE, priklausomai nuo oro uostų krypties. Seq2Seq modelis daugeliu atvejų tiksliausiai prognozavo vartotojų paieškas skirtingiems oro uostų deriniams.

Gytis Baravykas. Performance improvement of a travel planning website based on user behavior analysis and forecasting. Bachelor's / Master's Final Degree Project / Final Degree Project of Minor Studies / Professional Studies (choose one) / supervisor PhD. Agnė Paulauskaitė-Tarasevičienė; Informatics Faculty, Kaunas University of Technology.

Study field and area (study field group): Computer science, Informatics (B01).

Keywords: time series, forecasting, data analysis, travel search demand, user behavior patterns.

Kaunas, 2024. 59.

Summary

Predicting consumer choices has become an increasingly relevant task in the rapidly expanding realm of e-commerce. This issue is particularly significant due to the growing number of suppliers, the wide range of services available, and the evolving needs of consumers. This study aims to explore various artificial intelligence models for forecast travel search volumes. The literature review addresses the task of time series regression and the methods used, with a specific focus on solutions for forecasting flight demand. The analysis includes solutions from similar studies, encompassing the ARIMA statistical model, linear neural networks, long short-term memory networks (LSTM), and Sequence-to-Sequence (Seq2Seq) models. A consumer data analysis is conducted to identify user behavior patterns in travel searches. The travel search data characteristics discussed include multiple time dimensions, geo objects and encoding solutions for these features. Two approaches for predicting search volumes are examined: the first involves transforming data based on the flight departure date, and the second based on the flight search date. The study employs ARIMA, Prophet, linear multilayer neural networks, LSTM and Seq2Seq autoencoder models. Experiments are conducted in which the models are trained to forecast 1,7 and 30 time steps ahead for various airport combinations. The results for predictions based on the departure date indicate that the Seq2Seq model achieved the best performance with a 52.3% MAPE error on test data. The LSTM model had a prediction error of 59.7% MAPE, and the ARIMA model had an error of 56% MAPE. Experiments based on the search date showed that geographic objects are suitably encoded with coordinates. Results based on the search date demonstrated that single step predictions with neural network models were quite accurate, ranging from 22-25% MAPE. For multi-step forecasting tasks, LSTM and Seq2Seq models achieved better results compared to the linear multilayer neural network, with an average improvement of 10-30% MAPE, depending on the airport direction. The Seq2Seq model most accurately predicted user searches for the majority of airport routes.

Turinys

Lentelių sąrašas

Paveikslų sąrašas

Santrumpų ir terminų sąrašas

Santrumpos:

Doc. – docentas;

Lekt. – lektorius;

Prof. – profesorius.

Terminai:

LR - Tiesinė regresija

RF – Atsitiktinių medžių miškas

GBT – Gradientu stiprinti medžiai

NLP – Natūralios kalbos apdorojimas

HA – Istorinis vidurkis (angl. Historical average)

MA – judantis vidurkis (angl. Moving average)

ANN – dirbtinis neuroninis tinklas

RNN – Rekurentinis neuroninis tinklas

GRU - reguliuojami pasikartojantys vienetai

ETS – Klaida, Tendencija, Sezoniškumas

LSTM - ilgoji trumpalaikė atmintis

ARIMA – Auto regresinis apverstas judantis vidurkis

Įvadas

Internetas yra viena populiariausių vietų pirkti įvairias paslaugas ir prekes. Tiekėjai yra suinteresuoti įtraukti pirkėją ir sudaryti geriausias sąlygas užsakymo peržiūrai ir įvykdymui. Norint tai padaryti sklandžiai internetinėje erdvėje yra vis sudėtingiau. Kiekvienas gamintojas turi savo sandėlius ir kiekius, skirtingas kainas ir infrastruktūrą, o ir vartotojų kiekis internete tik daugėja.

Ne išimtis ir kelionių sektorius, kuris apima didelę dalį sferų: viešbučiai, būsto nuoma, skrydžiai, automobilių nuoma, transporto paslaugos, restoranai ir įvairios pramogos. Kiekviena rinka turi tūkstančius tiekėjų, dar daugiau paslaugų ir prekių. Vartotojui yra sunku palyginti ir išsirinkti, dėl ko atsiranda įvairūs kelionių paieškos įrankiai kaip „Momondo“, „Booking.com“ ar „Kayak“, kurie bendradarbiaudami su partneriais, surenka ir pateikia pasirinkimus vartotojams. Yra ne viena kompanija galinti pasiūlyti tuos pačius rezultatus ir pasirinkimus. Konkurencija yra stipri, dėl to svarbu klientams pateikti geriausią patirtį, kitu atveju klientai išeis pas konkurentus. Sklandžiai ir greitai pateikti rezultatus klientams tampa vis sunkiau dėl šių priežasčių:

- platus paslaugų pasirinkimas;
- augantis tiekėjų skaičius;
- vis didėjantis vartotojų kiekis;
- kintanti paslaugų paklausa ir pasiūla;
- nepastovūs vartotojų poreikiai.

Išvardintos priežastys sudaro sudėtingas sąlygas paslaugos optimizavimui, kadangi rezultatai negali būti statiški ilgą laiką, o nuolat kreiptis į tiekėjus dėl paprastos užklausos irgi ne išeitis. Vystantis rinkai atsiranda nauji partneriai, nauji duomenų šaltiniai ir rezultatų gavimas iš daugybės tiekėjų trunka vis ilgiau. Bet jeigu mes galėtume prognozuoti ko nori mūsų vartotojai ir pasiruošti iš anksto? Prognozuoti skrydžių paieškų kiekius yra žmogui sunkiai įgyvendinama užduotis, reikia išnagrinėti vartotojų elgseną, duomenų tendencijas, oro uostų pasirinkimus ir daug kitų duomenų. Statistiniai modeliai ir giliojo mokymo algoritmai jau daugelį metų naudojami analogiškose problemose: inventoriaus prognozė, keleivių kiekių prognozė ir kiti. Komercinės skrydžių paieškos yra mažai ištyrinėta sritis, turinti nestandartinius bruožus laiko eilučių uždaviniuose, tokius kaip: kelios laiko vertikalės, aukšto kardinalumo kategorinės vertės ir kintančio ilgio tarpai tarp duomenų rinkinių.

Darbo tikslas – ištirti įvairius dirbtinio intelekto ir statistinius modelius kelionių paieškų kiekio prognozavimui išskiriant ir įvertinant vartotojų elgsenos šabloninius bruožus ieškant kelionių.

Uždaviniai:

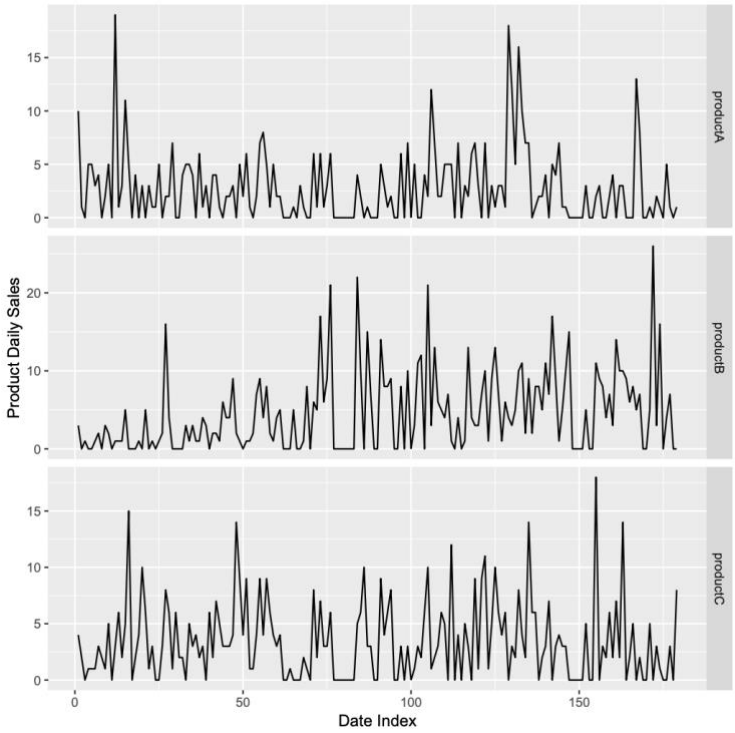
1. išanalizuoti dirbtinio intelekto ir statistinius modelius skirtus paieškos parametrų prognozavimui;
2. sudaryti duomenų rinkinius kaupiančius reprezentatyvią informaciją apie vartotojų paieškas ir atlikti vartotojų duomenų analizę;
3. ištirti esamų prognozavimo modelių galimybes siekiant prognozuoti kelionių paieškų skaičių;
4. sukurti sprendimus grįstus statistiniais ir dirbtinio intelekto modeliais siekiant prognozuoti paieškų kiekius pagal du parametrus įskaitant skrydžio išvykimo datą ir skrydžio paieškos datą;
5. atlikti eksperimentinius tyrimus varijuojant prognozavimo į priekį laiko masteliu;
6. įvertinti realizuotų metodų rezultatus, atlikti palyginamąją analizę ir patiekti įžvalgas.

1. Literatūros analizė

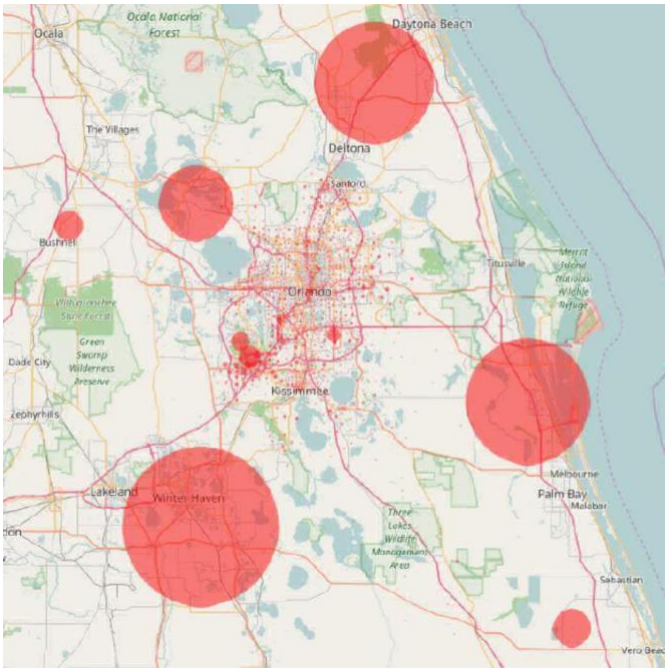
1.1. Laiko eilučių prognozavimo uždavinys

Laiko eilučių prognozės uždavinys ir jo variacijos yra viena pagrindinių mašininio mokymo sričių, kurią bando išspręsti ir pagerinti daugelis duomenų mokslo specialistų ir tyrėjų. Uždavinys yra prognozuoti būsimus įvykius, vertes ar taškus pagal istorinių duomenų pasitaikančias variacijas. Šie įvykiai pasitaiko dažniausiai tam tikrais laiko intervalais (kas valandą, kas dieną, kas mėnesį, kas metų ketvirtį). Uždavinys pasitaiko įvairiose pramonės srityse kaip finansų, prekybos ir produktų, logistikos, meteorologijos, energetikos bei daugelyje kitų sričių. Šiose srityse prognozuojami elementai kaip akcijų pakitimai, prekių kiekiai, transporto srautai, temperatūros pokyčiai. Prognozės leidžia geriau planuoti resursus. Šis uždavinys susilaukia didelio susidomėjimo, nes pastebėti sekas duomenyse nėra paprasta dėl besikeičiančių laiko eilučių variacijų (sezoniškumas, nepastovumas, anomalijos).

Į laiko eilučių prognozės sritį patenka regresijos skiltis, kuri prognozuoja būtent skaitines vertes pasireiškiančias laiko eilutėse (žr. 1 pav., 2 pav.). Skilties principas yra aptikti asociacijas tarp įvairių verčių laiko eilutėse ir atrasti kaip pokyčiai veikia galutinį rezultatą. Tai gali būti prekių pardavimų priklausomybė nuo parduotuvių kiekio ar kurjerių darbo laiko. Šitos asociacijos padeda modeliams geriau nuspėti pokyčius ateityje.



1 pav. Laiko eilučių regresinių verčių pavyzdys. Produktų kiekiai dienų indeksų metų [1].



2 pav. Regresijos duomenų pavyzdys. Vairuotojų kelionių kiekiai tam tikram regione [2].

Esant dideliems duomenų kiekiams, prognozavimo užduotis tampa sunkiai įgyvendinama, dėl to prognozėms generuoti taikomi įvairūs statistiniai matematiniai metodai, dirbtiniai neuroniniai tinklai ir jų variacijos.

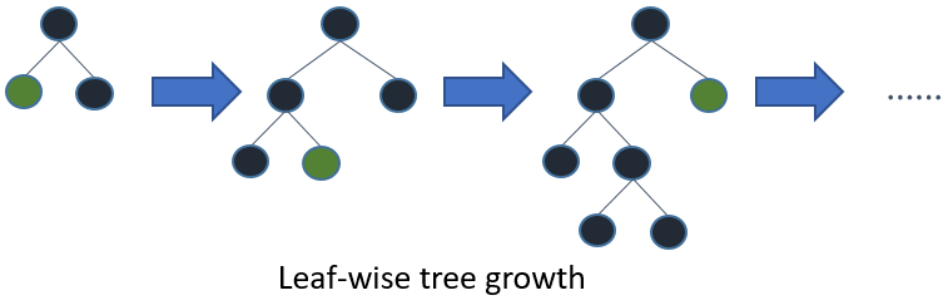
Per daugelį metų daryta daug eksperimentų su statistiniais modeliais, kurių principas – transformuoti duomenis, siekiant išgauti reikiamą atsakymą. Tokie metodai kaip ETS, ARIMA, Nai‘ve ir AR [3]. Šie statistiniai modeliai turi labai gerą tikslumą dėl kurio, net ir naujesni tiriamieji darbai naudoja juos kaip kontrolinę grupę lyginant su sudėtingesniais neuroninių tinklų plėtiniais [2]. Industrijose dar plačiai naudojami statistiniai metodai dėl jų patikimumo, paprastumo, greičio ir tikslumo [4], tačiau pastebėta, kad šiems metodams neretai reikia nemažai duomenų apsimokinimui norint išgauti didesnį tikslumą. Statistinių modelių pranašumas dingsta sudėtingėjant uždaviniams, atsirandant įvairesnių duomenų savybių ir staigių pakitimų, kadangi pagal istorinius duomenis sprendimus priimantys statistiniai modeliai nesugeba interpretuoti n-mačių duomenų savybių [3]. Tokiems modeliams taip pat reikia specialiai paruošti duomenis arba skaidyti duomenis į mažesnes kategorijas [4], dėl to gali stipriai prailgėti užduoties įgyvendinimas.

Šalia statistinių metodų, standartiniai mašininio mokymo metodai vis dar aktualūs spendžiant įvairias laiko prognozės eilučių problemas. Tai tokie standartiniai modeliai, kaip paprasti dirbtiniai neuroniniai tinklai, sprendimų medžiai, atsitiktinis medžių miškas, atramos vektorių mašina ir tiesinė regresija. Šie baziniai modeliai yra pasiekę gerų rezultatų mašininio mokymo sprendimuose. Tiesinė regresija yra vienas paprasčiausių ir nesunkiai suprantamų modelių, naudojamų ateities verčių prognozėms [4]. Šis tiesinės regresijos paprastumas, tikslumas ir aiškumas vidiniuose skaičiavimuose yra viena pagrindinių priežasčių dėl ko metodas dažnai imamas kaip bazinis modelis lyginimui su naujomis sistemomis, kaip transformeriai ar LSTM [2,5,6].

Ne išimtis ir atsitiktinių medžių miškas, kuris 2016 metų tyrime, patikimai prognozavo vieną sunkiausių duomenų rinkinių – NASDAQ akcijos pardavimai. Pasiekdamas, pagal rezultatus, tikslumą nuo 92% iki 94% prognozuojant 3 mėnesių vertes bendroje akcijų skiltyje [7]. Kadangi standartiniai modeliai yra ypač populiarūs sprendžiant mašininio mokymo problemas, tyrėjai vis

tobulina šiuos algoritmus pridėdami variacijų [8,9]. Laiko eilučių sekoms ir priklausomybėms aptikti atsitiktiniai medžiai turi sukurti labai sudėtingą modelį, kas neretai priveda prie modelio persimokymo ir nestabilumo, kada, bet kokia nauja variacija įtakoja kitaip augti atsitiktiniam medžiui [7]. Šis sudėtingas atsitiktinio medžio augimas žymiai prailgina treniravimo ir prognozavimo laiką, ypatingai pasireiškia esant vis didesniems duomenų kiekiams. Šias problemas išsprendžia naujesnės atsitiktinių medžių miško versijos kaip:

1. Gradientu sustiprinti medžiai (angl. Gradient Boosting Trees – GBT) – šio metodo privalumas, kad sprendimų medžiai yra sukuriami vienas po kito. Šis principas leidžia medžiams mokytis ir atitaisyti klaidas iš prieš tai buvusių. Toks modelis turi privalumų kaip aukštesnis duomenų prognozavimo tikslumas ir atsparumas anomalijoms.
2. Lengvasis gradiento stiprinimo mechanizmas (angl. Light Gradient Boosted Mechanism – LGBM) – šio metodo privalumas yra atšakos skaidymas lapų lygmenyje gilinant medį vertikaliai (žr. 3 pav.) vietoj standartinio horizontalaus skaidymo. Toks metodas yra greitesnis ir tausojantis resursus kaip kompiuterinė procesoriaus atmintis.



3 pav. LGBM medžio auginimo principas [10].

Šios naujos atsitiktinių medžių miškų variacijos (GBT, LGBM) atsparesnės, greičiau generuojamos, dėl to neblogai konkuruoja su naujomis neuroninių tinklų variacijomis [11,12].

Kiekvienas standartinių mašininio mokymo metodų rodo neblogus rezultatus susiduriant su laikų eilučių prognozės užduotimi, dėl ko kyla klausimas, kodėl neapjungus jų į vieną grupę, kurios bendra prognozė būtų dar tikslesnė. Šis darbas [13] tyrinėjo kaip tradicinių mašininio algoritmų „ansamblis“ prognozuoja elektroninės parduotuvės prekių paklausas. Darbe ištyrinėti pavieniai algoritmai: RF, LR, DT, GBT ir įvairios sudurtinės paminėtų algoritmų „ansamblių“ variacijos.

Model	RMSE
DT	2.200
GBT	2.299
RF	2.120
LR	1.910

Model	RMSE
DT	1.928
GBT	1.918
RF	1.865
LR	2.708
SG(LR)	1.864

Model	RMSE
GBT+DT	1,955
GBT+LR	1,957
LR+DT	1,962
DT+RF	1,963
GBT+RF	1,870
LR+RF	1,927

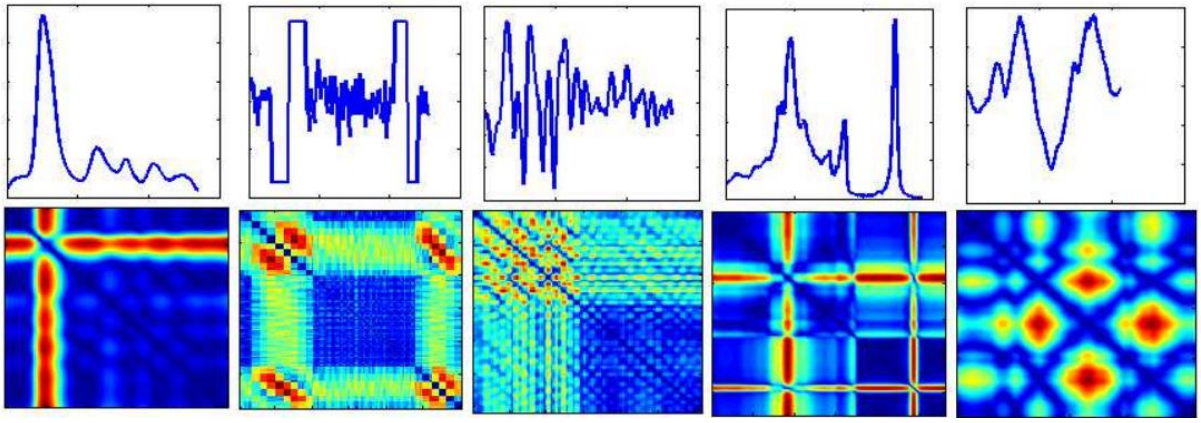
Model	RMSE
DT+RF+GBT	1.894
RF+DT+LR	1.909
GBT+LR+DT	2.011
LR+RF+GBT	1.864

4 pav. Tradicinių mašininio mokymo algoritmų ir jų "ansamblių" RMSE rezultatai [13].

Kaip matyti iš tyrimo rezultatų (žr. 4 pav.) kartu sugrupuoti metodai pasiekia tikslesnes duomenų prognozes. Straipsnis neatsižvelgia kiek kompleksiškumo sukuria kelių modelių derinimas kartu, treniravimas ar papildomi resursai, šie faktoriai gali užgožti tikslumo padidėjimo naudą.

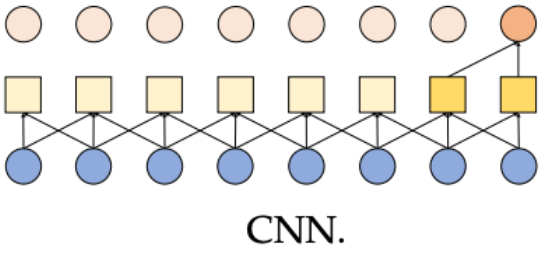
Atradimai giliųjų neuroninių tinklų srityje atvėrė naujas galimybes ir laiko eilučių uždaviniuose. Gilieji neuroniniai tinklai išmoksta netiesines priklausomybes, dėl to gali būti pritaikomi vietose kur statistinių metodų negalima panaudoti. Giliojo neuroninio tinklo esminė savybė – apsimokinti ir išgauti daugiaskalę informaciją iš didelių duomenų rinkinių.

Laiko eilučių prognozė yra vienas aktualiausių uždavinių mašininio mokymo srityje, dėl to nemažai tyrėjų bando pernaudoti modelius iš kitų uždavinių, tokių kaip kompiuterinė rega (angl. computer vision). Kompiuterinės regos srityje plačiai taikomi konvoliuciniai neuroniniai tinklai (CNN) buvo bandomi pritaikyti ir laiko eilučių uždaviniams spręsti. Norint naudoti CNN laiko eilučių užduotims pirmiausia reikia šiuos duomenis paversti į vaizdinę formą arba kitais žodžiais, dvimatę matricą [14]. Pavertus duomenis į vaizdinį formatą (žr. 5 pav.) CNN gali sugrupuoti kartu artimas datas ir jų kiekius [15].

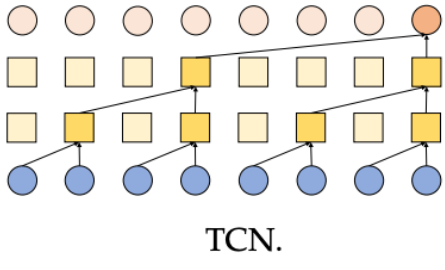


5 pav. Įvairių laiko eilučių duomenų rinkinių užkodavimas vaizdiniu formatu [15] .

Generuoti vaizdus iš didelio savybių kiekio gali tapti sudėtingu uždaviniu, tačiau yra sukurtos CNN variacijos specialiai pritaikytos laiko eilučių analizei. Į konvoliucinius sluoksnius įterpiami kintamo dydžio filtrai, kad būtų pastebima įvairių skalių laike esanti informacija [16]. Viena iš sukurtų variacijų specialiai pritaikytų laiko eilučių uždaviniams yra laiko konvoliuciniai tinklai (TCN). Skirtingai nuo įprastų konvoliucinių neuroninių tinklų (žr. 6 pav.), laiko konvoliuciniai tinklai (žr. 7 pav.) naudoja pilną konvoliucinį tinklą, kuriame visi sluoksniai yra tokio pat ilgio, ir naudoja priežastinės konvoliucijas, kurios nepraranda informacijos apdorojant ateities ir praeities duomenis.



6 pav. Paprasto konvoliucinio tinklo schema [17]



7 pav. Paprasto laiko konvoliucinio tinklo schema [17]

Variacijos TCN metodo išbandytos ant įvairių laiko eilučių duomenų rinkinių: eismo, elektros sąnaudų ir automobilių atsarginių detalių paklausos. TCN parodė aukštą efektyvumą, paprastumą ir tikslumą sekos uždaviniuose palyginus su dažnai naudojamais mašininio mokymo modeliais kaip

RNN, LGBM, AR ir ETS [12]. Vienas iš svarbiausių atradimų laiko eilučių uždaviniuose tai rekurentiniai neuroniniai tinklai (RNN). Skirtingai negu tiesinės sekos neuroniniai tinklai (FFNN), rekurentiniai neuroniniai tinklai sukonstruoti su paslėptomis būsenomis, kurios leidžia saugoti ir panaudoti informaciją iš ankstesnių pasikartojančių įvykių [18]. RNN pasižymi funkcionalumu, kurios aktualios sprendžiant pasikartojančių sekų uždavinius:

- Pasikartojančių sekų priklausomybių konstravimas.

  Modelis sugeba sukurti asociacijas ir atsižvelgti į duomenų eiliškumą laiko spektre. Tokiu būdu modelis pastebi ilgalaikes priklausomybes ir išmoksta šablonus, kurie kinta ir vystosi pagal laiko pokytį. Išlaikant vidines būsenas modelis kaupia informaciją iš ankstesnių laiko žingsnių, kurie vėliau panaudojami formuoti prognozėms tolimesnėje modelio stadijoje.

- Kintančio ilgio įvesčių apdorojimas

  Galimybė apdoroti įvairių ilgių paduodamus duomenis leidžia RNN geriau prisitaikyti prie laiko eilučių duomenų, kurie gali turėti skirtingus sekų ilgius.

Išvadintos savybės ir jų aktualumas šio tipo uždaviniuose yra priežastys dėl ko šis modelis arba jo variacijos naudojamos daugelyje sričių kaip signalų analizė, finansinių akcijų prognozės ir meteorologijos sąlygų nuspėjimas [18].

Dėl rekurentinių neuroninių tinklų populiarumo sekų uždaviniuose, buvo pastebėtos kelios problemos:

- Nykstantis arba sprogstantis gradientas - apmokymo metu gradientas arba tiksliau jo vertės vis mažėja arba sparčiai didėja keliaudamos per neuronų sluoksnius. Toks gradientas, grįžus prie pradinių neuronų sluoksnių nebeturi poveikio ant pradinių neuronų svorių. Svoriai lieka beveik nepakitę arba per daug pakitę. Nykstant ar sprogstant gradientui, rekurentiniam tinklui tampa sudėtinga išmokti ilgalaikes duomenų priklausomybes. Šis apribojimas turi įtakos RNN gebėjimui fiksuoti ir išlaikyti informaciją ilgesnėse sekose.

- Limituota atmintis - tradicinių RNN atminties talpa yra ribota, todėl esant dideliam duomenų kiekiui ar ilgoms duomenų sekos, sunkiai atsimenama informacija iš tolimiausių žingsnių. Šitą reiškinį apsunkina nestabilus gradiento irimas, atsiranda šališkumas, labiau pastebintis naujesnes duomenų įvestis. Ši problema, įtakoja, kad RNN taps sudėtinga išlaikyti priklausomybės ryšius tarp laiko sekų.

Šias problemas išsprendžia RNN modifikuotos variacijos kaip: ilgos-trumpos atminties (LSTM) ir vartiniai rekurentiniai vienetai (GRU). LSTM architektūra išsprendžia šias problemas įkomponuodama atminties blokus ir kontroliuojamą vartų sistemą. Ši sistema sudaryta iš trijų vartų: įvesties, užmaršos ir išvesties. Įvesties vartai valgo paduodamų duomenų srautą, kurie patenka į atminties blokus. Užmaršos vartai valdo praeities informacijos saugojimą. Išvesties vartai atsakingi už rezultatus, kurie reguliuojami pagal įvesties informaciją ir atminties blokų būsenas. Tokia architektūra tausoja modelio atmintį ir sutvarko gradiento vertes. Šie patobulinimai nuo standartinio RNN, padeda modeliui užfiksuoti ir išlaikyti ilgesnių sekų informaciją [19].

GRU tai dar viena RNN modelio variacija [18], turinti paprastesnę architektūrą negu LSTM, tačiau pasiekia panašų našumą. GRU architektūroje atminties blokas ir paslėpta būsena sujungiami į vienus atnaujinimo vartus, o informacijos srautui valdyti sukuriami atstatymo vartai. Atnaujinimo vartai reguliuoja kiek ankstesnių būsenų reikia išsaugoti ir kiek naujos informacijos įtraukti, o atstatymo

vartai atsakingi už informacijos kiekio pamiršimą. Tokia architektūra turi mažiau parametrų, kas skaičiavimo požiūriu yra efektyviau.

Apžvelgus teorines šių dviejų RNN variacijų dalis, galima būtų teigti, kad GRU algoritmas turėtų būti plačiau naudojamas ir efektyvesnis prieš LSTM, tačiau praktika yra kiek kitokia. Išanalizavus įvairius mokslinius straipsnius matoma, kad LSTM modelis yra labiau patrauklesnis negu GRU laiko eilučių sprendimuose [1,2,4,20,21]. Laiko eilučių uždaviniuose giliųjų neuroninių tinklų modeliai kaip LSTM rodo geresnius rezultatus negu statistiniai metodai ar paprasti neuroniniai tinklai. Šis darbas [4] ištyrinėjo kaip LSTM ir jo variacijos prognozuoja keleivių pavėžėjimo paklausas palyginus su statistiniais modeliais ir paprastais neuroniniais tinklais. Kaip matome iš (žr. 8 pav.) FCL-Net, kuris yra sudurtinis konvoliucinis LSTM tinklas, turi beveik dvigubai geresnį tikslumą negu kiti metodai, dėl to, kad standartiniai statistiniai modeliai nėra pritaikyti įsisavinti n-mačių savybių [4].

Model	RMSE	R^2	MAE
HA	0.0378	0.736	0.0192
MA	0.0511	0.518	0.0260
ARIMA	0.0345	0.780	0.0178
ANN	0.0331	0.798	0.0194
LSTM	0.0322	0.808	0.0181
Conv-LSTM (with only demand intensity)	0.0318	0.813	0.0176
FCL-Net (with full variables)	0.0156	0.820	0.0090
FCL-Net (with selected variables)	0.0157	0.819	0.0091

8 pav. Prognozavimo tikslumo palyginimas tarp įvairių metodų [4].

Vienas įdomesnių modelių pasiūlytų spręsti sekos uždaviniams yra Seq2Seq [22], kuris turi skirtingą struktūrą palyginus su kitomis neuroninių tinklų variacijomis. Architektūra pasižymi tuo, kad Seq2Seq modelį sudaro du pagrindiniai komponentai: užkodavimo ir dekodavimo blokai.

Kodavimo blokas apdoroja įvesties duomenis, pavyzdžiui, istorines laiko eilutes, ir paverčia juos fiksuoto ilgio vektorine reprezentacija dar vadinamu turinio vektoriumi. Turinio vektorius atspindi pagrindines įvesties duomenų sekas. Užkodavimo komponentas gali būti, bet kokia RNN modelio variacija [22].

Dekodavimo įrenginys paima konteksto vektorių kaip įvestį ir sugeneruoja išvesties seką, dar kitaip tariant rezultatą, pavyzdžiui, prognozuojamų duomenų seką. Dekodavimo komponentas yra treniruojamas generuoti kuo tikslesnes prognozes remiantis konteksto vektoriumi ir ankstesniais sugeneruotais rezultatais. Dekodavimo komponentas taip pat gali būti RNN modelio variacija.

Vienas naujausių atradimų dirbtinio intelekto skiltyje – transformerių architektūros. Transformerio architektūra pagrįsta savaiminio dėmesio mechanizmų konceptu. Šis konceptas nurodo, kad modelis sugeba apsimokinti prognozių generavimo metu, sutelkus dėmesį į skirtingas įvesties sekos dalis. Transformerio architektūrą sudaro užkodavimo ir dekodavimo komponentai, panašiai kaip Seq2Seq modelis, tačiau esminis skirtumas – dėmesio mechanizmas. Transformerio dėmesio mechanizmo dėka efektyviau užfiksuojamos ilgalaikės įvesties sekos priklausomybės,

palygus su tradiciniais rekurentiniais neuroniniais tinklais (RNN). Tai pasiekiama vienu metu atkreipiant dėmesį į visas įvesties sekos pozicijas ir jų svarbą, vietoj nuoseklaus įvesties apdorojimo. Transformeriai yra pasiekę puikius rezultatus kituose mašininio mokymo srityse kaip kompiuterinė rega, natūralios kalbos procesai (NLP), tačiau tyrimai rodo paprastesnius rezultatus laiko eilučių prognozės atžvilgiu. Atliktas tyrimas [5], kuriame naujausi transformerių modeliai pritaikyti laiko eilučių prognozėms buvo lyginami prieš paprastas tiesinės regresijos variacijas, tokias kaip tiesinė variacija su papildomais duomenų skaidymais į tendenciją ir likutines vertes. Šiame tyrime buvo naudojami įvairus elektros, eismo, meteorologiniai duomenų rinkiniai.

Methods		IMP.	Linear*		NLinear*		DLinear*		FEDformer		Autoformer		Informer		Pyraformer*		LogTrans		*Repeat**	
Metric		MSE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	96	27.40%	**0.140**	**0.237**	0.141	**0.237**	**0.140**	**0.237**	0.193	0.308	0.201	0.317	0.274	0.368	0.386	0.449	0.258	0.357	1.588	0.946
	192	23.88%	**0.153**	0.250	0.154	**0.248**	**0.153**	0.249	0.201	0.315	0.222	0.334	0.296	0.386	0.386	0.443	0.266	0.368	1.595	0.950
	336	21.02%	**0.169**	0.268	0.171	**0.265**	**0.169**	0.267	0.214	0.329	0.231	0.338	0.300	0.394	0.378	0.443	0.280	0.380	1.617	0.961
	720	17.47%	**0.203**	0.301	0.210	**0.297**	**0.203**	0.301	0.246	0.355	0.254	0.361	0.373	0.439	0.376	0.445	0.283	0.376	1.647	0.975
Exchange	96	45.27%	0.082	0.207	0.089	0.208	**0.081**	0.203	0.148	0.278	0.197	0.323	0.847	0.752	0.376	1.105	0.968	0.812	**0.081**	**0.196**
	192	42.06%	0.167	0.304	0.180	0.300	**0.157**	0.293	0.271	0.380	0.300	0.369	1.204	0.895	1.748	1.151	1.040	0.851	0.167	**0.289**
	336	33.69%	0.328	0.432	0.331	0.415	**0.305**	0.414	0.460	0.500	0.509	0.524	1.672	1.036	1.874	1.172	1.659	1.081	**0.305**	**0.396**
	720	46.19%	0.964	0.750	1.033	0.780	**0.643**	**0.601**	1.195	0.841	1.447	0.941	2.478	1.310	1.943	1.206	1.941	1.127	0.823	0.681
Traffic	96	30.15%	**0.410**	0.282	**0.410**	**0.279**	**0.410**	0.282	0.587	0.366	0.613	0.388	0.719	0.391	2.085	0.468	0.684	0.384	2.723	1.079
	192	29.96%	**0.423**	0.287	**0.423**	**0.284**	**0.423**	0.287	0.604	0.373	0.616	0.382	0.696	0.379	0.867	0.467	0.685	0.390	2.756	1.087
	336	29.95%	0.436	0.295	**0.435**	**0.290**	0.436	0.296	0.621	0.383	0.622	0.337	0.777	0.420	0.869	0.469	0.734	0.408	2.791	1.095
	720	25.87%	0.466	0.315	**0.464**	**0.307**	0.466	0.315	0.626	0.382	0.660	0.408	0.864	0.472	0.881	0.473	0.717	0.396	2.811	1.097
Weather	96	18.89%	**0.176**	0.236	0.182	**0.232**	**0.176**	0.237	0.217	0.296	0.266	0.336	0.300	0.384	0.896	0.556	0.458	0.490	0.259	0.254
	192	21.01%	**0.218**	0.276	0.225	**0.269**	0.220	0.282	0.276	0.336	0.307	0.367	0.598	0.544	0.622	0.624	0.658	0.589	0.309	0.292
	336	22.71%	**0.262**	0.312	0.271	**0.301**	0.265	0.319	0.339	0.380	0.359	0.395	0.578	0.523	0.739	0.753	0.797	0.652	0.377	0.338
	720	19.85%	0.326	0.365	0.338	**0.348**	**0.323**	0.362	0.403	0.428	0.419	0.428	1.059	0.741	1.004	0.934	0.869	0.675	0.465	0.394
ILI	24	47.86%	1.947	0.985	**1.683**	**0.858**	2.215	1.081	3.228	1.260	3.483	1.287	5.764	1.677	1.420	2.012	4.480	1.444	6.587	1.701
	36	36.43%	2.182	1.036	**1.703**	**0.859**	1.963	0.963	2.679	1.080	3.103	1.148	4.755	1.467	7.394	2.031	4.799	1.467	7.130	1.884
	48	34.43%	2.256	1.060	**1.719**	**0.884**	2.130	1.024	2.622	1.078	2.669	1.085	4.763	1.469	7.551	2.057	4.800	1.468	6.575	1.798
	60	34.33%	2.390	1.104	**1.819**	**0.917**	2.368	1.096	2.857	1.157	2.770	1.125	5.264	1.564	7.662	2.100	5.278	1.560	5.893	1.677
ETTh1	96	0.80%	0.375	0.397	**0.374**	**0.394**	0.375	0.399	0.376	0.419	0.449	0.459	0.865	0.713	0.664	0.612	0.878	0.740	1.295	0.713
	192	3.57%	0.418	0.429	0.408	**0.415**	**0.405**	0.416	0.420	0.448	0.500	0.482	1.008	0.792	0.790	0.681	1.037	0.824	1.325	0.733
	336	6.54%	0.479	0.476	**0.429**	**0.427**	0.439	0.443	0.459	0.465	0.521	0.496	1.107	0.809	0.891	0.738	1.238	0.932	1.323	0.744
	720	13.04%	0.624	0.592	**0.440**	**0.453**	0.472	0.490	0.506	0.507	0.514	0.512	1.181	0.865	0.963	0.782	1.135	0.852	1.339	0.756
ETTh2	96	19.94%	0.288	0.352	**0.277**	**0.338**	0.289	0.353	0.346	0.388	0.358	0.397	3.755	1.525	0.645	0.597	2.116	1.197	0.432	0.422
	192	19.81%	0.377	0.413	**0.344**	**0.381**	0.383	0.418	0.429	0.439	0.456	0.452	5.602	1.931	0.788	0.683	4.315	1.635	0.534	0.473
	336	25.93%	0.452	0.461	**0.357**	**0.400**	0.448	0.465	0.496	0.487	0.482	0.486	4.721	1.835	0.907	0.747	1.124	1.604	0.591	0.508
	720	14.25%	0.698	0.595	**0.394**	**0.436**	0.605	0.551	0.463	0.474	0.515	0.511	3.647	1.625	0.963	0.783	3.188	1.540	0.588	0.517
ETTm1	96	21.10%	0.308	0.352	0.306	0.348	**0.299**	**0.343**	0.379	0.419	0.505	0.475	0.672	0.571	0.543	0.510	0.600	0.546	1.214	0.665
	192	21.36%	0.340	0.369	0.349	0.375	**0.335**	**0.365**	0.426	0.441	0.553	0.496	0.795	0.669	0.557	0.537	0.837	0.700	1.261	0.690
	336	17.07%	0.376	0.393	0.375	0.388	**0.369**	**0.386**	0.445	0.459	0.621	0.537	1.212	0.871	0.754	0.655	1.124	0.832	1.283	0.707
	720	21.73%	0.440	0.435	0.433	0.422	**0.425**	**0.421**	0.543	0.490	0.671	0.561	1.166	0.823	0.908	0.724	1.153	0.820	1.319	0.729
ETTm2	96	17.73%	0.168	0.262	**0.167**	**0.255**	**0.167**	0.260	0.203	0.287	0.255	0.339	0.365	0.453	0.435	0.507	0.768	0.642	0.266	0.328
	192	17.84%	0.232	0.308	**0.221**	**0.293**	0.224	0.303	0.269	0.328	0.281	0.340	0.533	0.563	0.730	0.673	0.989	0.757	0.340	0.371
	336	15.69%	0.320	0.373	**0.274**	**0.327**	0.281	0.342	0.325	0.366	0.339	0.372	1.363	0.887	1.201	0.845	1.334	0.872	0.412	0.410
	720	12.58%	0.413	0.435	**0.368**	**0.384**	0.397	0.421	0.421	0.415	0.433	0.432	3.379	1.338	3.625	1.451	3.048	1.328	0.521	0.465

9 pav. Transformerių palyginimas prieš tiesinės regresijos variantus ant įvairių duomenų rinkinių [5].

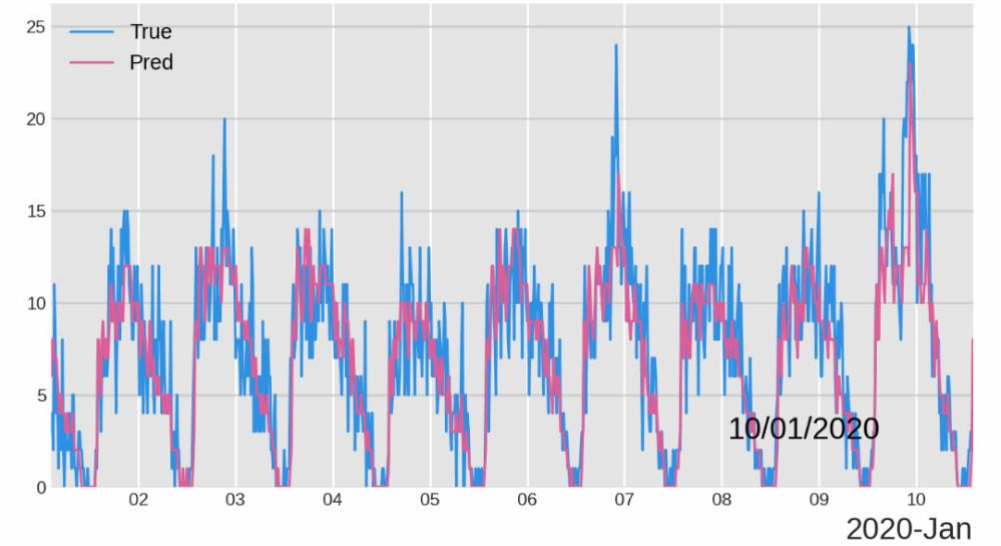
Rezultatai pateikti (žr. 9 pav.) rodo, kad transformeriai gali būti ne pats geriausias sprendimas laiko eilučių prognozavimui. Tiesa, peržiūrėjus pateiktus duomenų rinkinius ir transformerių parametrų kiekį (žr. 10 pav.) kyla teorijų į kurias autoriai neatsižvelgė. Tokių kaip: galimai duomenų rinkiniai yra per paprasti, nesudaro kompleksiškų sekų, kurias reikėtų apdoroti. Transformeriai turi labai didelius treniruojamų parametrų kiekius, taigi treniruojant juos ant labai paprastų duomenų, modelis gali persitreniruoti ir prastai prognuozuoti ateities vertes.

Method	MACs	Parameter	Time	Memory
DLinear	**0.04G**	**139.7K**	**0.4ms**	**687MiB**
Transformer×	4.03G	13.61M	26.8ms	6091MiB
Informer	3.93G	14.39M	49.3ms	3869MiB
Autoformer	4.41G	14.91M	164.1ms	7607MiB
Pyraformer	0.80G	241.4M*	3.4ms	7017MiB
FEDformer	4.41G	20.68M	40.5ms	4143MiB

10 pav. Modelių treniruojamų parametrų kiekis, prognozei generuoti reikalingas laikas ir užimamas atminties kiekis.

1.2. Skrydžių paklausos prognozavimo uždavinys

Aviacijos tematika mašininio mokymo srityje yra gana aktuali. Daugelis oro linijų ir paieškos įrankių, tiria dirbtinio intelekto pritaikymo būdus aviacijos srityse. Skrydžių paklausos prognozavimo uždavinys yra variacija laiko eilučių uždavinio. Panašūs straipsniai, kada bandyta prognozuoti oro uostų keleivių srautus [2,20]. Viename moksliniame tyrime [2] eksperimentams naudojami Las Vegaso oro uosto keleivių duomenys, siekiant prognozuoti keleivių srautus. Straipsnyje išbandyti įvairūs modeliai, išskirtinis Seq2Seq modelis uždavinyje pasirodė 60% geriau negu AR modeliai. Straipsnyje aprašytas Seq2Seq dėmesio modelis su maža klaida prognozuoja valandinius keleivių srauto pokyčius (žr. 11 pav.).



11 pav. Seq2Seq dėmesio modelis. LAS oro uosto skrydžių srauto prognozės rezultatai [2].

Iš rezultatų matome, kad Seq2Seq dėmesio modelis gana tiksliai aptinka pakitimus ir anomalijas, kurios gali reikšti specialių įvykių dienas. [20] Straipsnis tyrinėjo kelių oro maršrutų atvykstančių skrydžių kiekių prognozavimą, kuriame LSTM ir grafinė konvoliucinė LSTM versija generavo geriausias prognozes. Skrydžių kiekių nukėlimo uždavinyje ištirti tradiciniai mašininio mokymo algoritmai tokie kaip SVM, GBM, LGBM, RF ir variacijos, tarp kurių LGBM turėjo vieną geriausių prognozavimo tikslumą [23]. Tyrime pasirinkti keli populiarūs skrydžių maršrutai. Dar vienas straipsnis [24], tyrinėjo atvykstančių išvykstančių keleivių srautus, skirtingiems miestų deriniams. Tyrime lyginamas autorių sukurtas ATFPNet modelis,sudarytas iš GRU blokų, prieš GRU, GCN ir ANN modelius. Rezultatuose ATFPNet modelio prognozavimo tikslumas vos 1% aukštesnis negu standartinio GRU. Rasta publikacija apie skrydžių trajektorijos kiekius pagal išvykimo/atvykimo laiką [25]. Moksliniame darbe yra modeliuojamos skrydžių trajektorijos tarp oro uostų, grupuojami skrydžiai pagal išvykimo ir atvykimo laikus. Skrydžių trajektorijų kiekius, susikirtimo taškuose prognozavo [26]. Tyrime siūlomas autorių kurtas tranformerio tipo modelis, lyginamas prieš daugiasluoksnį perceptroną. Kiek skrydžių nukrypsta nuo pradinio išvykimo laiko [27]. Vienas panašiausių straipsnių [28], kuriame tiriamos galimybės prognozuoti oro uostų keleivių kiekius didžiausiuose oro uostuose pasaulyje nuo 2008 metų. Naudojami statistinis algoritmas VAR ir jo variacijos, duomenys turi tik vieną nenutrūkstančią laiko dimensiją. Kiekvienam oro uostui treniruojamas atskiras modelis. Tačiau, nei vienas nagrinėtas straipsnis neturi sprendimo didesniam oro uostų kiekiui, nenagrinėja laiko eilučių duomenų turinčių 2 laiko dimensijas.

2. Kelionių paieškos informacijos duomenų rinkiniai

Praktinis uždavinys yra prognozuoti paieškos kiekius oro uostų, išvykimo ir atvykimo datų kombinacijoms. Modelis turėtų mokyti ir pastebėti vartotojų elgsenos bruožus nulemiančius paieškų kiekį, tokius kaip: kiek dienų į priekį vartotojai dažniausiai žiūri, kokiam laikotarpiui, kokiu metų laiku, susidomėjimas specifinėmis datomis – gal koncertas ar įvykis.

2.1. Duomenų rinkiniai

Duomenų rinkinys sudarytas iš „KAYAK“ internetinės kelionių paieškos įrankio vartotojų skrydžių paieškų. Šios paieškos turi informaciją apie paieškos dieną, pradinį ir galutinį oro uostus, skrydžio išvykimo ir atvykimo datas.

2.2. Duomenų analizė ir paruošimas

Duomenų rinkinį sudaro 2021, 2022 ir 2023 metų vartotojų paieškos. Paieškos svetainė bendradarbiauja su įvairiomis oro linijomis, kurios bendrai apima apie 4 tūkstančių unikalių oro uostų ir maršrutus tarp jų.

paieškos data	išvykimo oro uostas	atvykimo oro uostas	išvykimo data	atvykimo data	paieškų kiekis
2021-01-01	IAH	LAX	2021-01-08	2021-01-10	23
2021-01-01	STI	JFK	2021-01-01	NaN	17
2021-01-01	MIA	DXB	2021-01-15	2021-01-22	12
2021-01-01	MIA	PHL	2021-01-05	NaN	15
2021-01-01	OTP	TSR	2021-01-01	NaN	12

12 pav. Duomenų pavyzdys.

Duomenų pavyzdyje (žr. 12 pav.) matomos laiko eilučių vertės savaime yra nenaudingos, kadangi statistiniai ir mašininio mokymo modeliai savaime neapdoroja datų ir tekstinių verčių. Šitas vertes reikia konvertuoti į skaitines, kad mašininio mokymo modeliai sugebėtų jas apdoroti. Yra mašininio mokymo algoritmų, kurie palaiko tekstines vertes, suskaitydami jas į kategorines vertes. Vienas iš jų LightGBM, kuris yra atsitiktinio miško variacija. Šis algoritmas pritaiko Fisher metodą [29] išgaunant homogeniškumą tarp kategorinių verčių.

```
išvykimo oro uostas           object
atvykimo oro uostas           object
išvykimo data         datetime64[ns]
atvykimo data         datetime64[ns]
paieškų kiekis                 int64
```

13 pav. Duomenų tipas kiekvienam stulpeliui.

Oro uostai yra tekstinės, aukšto kardinalumo vertės (žr. 13 pav.), tai sudaro problemas naudojant tokius užkodavimo metodus kaip „One-Hot encoding“, kuris sukuria per daug parametrų, kurie apsunkina teisingą modelių treniravimą [30]. Šie tyrimai [30,31] apžvelgia įvairias metodologijas

kaip užkoduoti aukšto kardinalumo kategorines savybes, tokias kaip n-gramos, šifravimas (hashing), dažnio užkodavimas, one-hot, label, sąsajų apsimokinimas (embedings), taip pat ir natūralios kalbos procesuose naudojami užkodavimo procesai kaip žodžių maišas, žodžio vektorinė reprezentacija (word2vector). NLP mašininio mokymo modeliai apsimokina sąsajas tarp žodžių, tai reiškia, kad modelio svorių matricoje giminingi žodžiai yra arčiau vienas kito, tačiau oro uostai neturi žodinio ryšio tarpusavyje, dėl ko šitie metodai yra netinkami. Kadangi oro uostai yra geografinės prasmės savybės, pavertus oro uostą į ilgumos ir platumos koordinates, mašininio mokymo modeliai turėtų neblogai aptikti sąsajas tarp verčių.

Datos savaime kaip tekstinės vertės neturi jokios naudingos informacijos, kurią galėtų apsimokinti algoritmai. Iš esamų datų reikia sugeneruoti savybes, kurios apibrėžia vartojo mąstymą ir elgseną. Iš datos ištraukiama informacija – metai, mėnuo, diena, savaitės diena, metų diena, metų savaitė, ketvirtis. Svarbūs dienų skirtumai tarp datų arba kitaip tariant deltos. Šios savybės modeliui suteikia galimybę lengviau pamatyti ilgalaikes tendencijas ir sezoniškumus.

Išgautas savybes galima naudoti modeliui apmokinti, tačiau daugelis šių savybių yra pasikartojančios, kitaip tariant ciklinės. Šis tyrimas [32] aptaria pakartotinių verčių užkodavimą sinuso ir kosinuso signalo formatu, tokiu būdu sumažinant laiko erdvės skirtumus tart sekmadienio ir pirmadienio arba tarp 23 valandos ir 1 valandos. Tyrimo rezultatai ir išvados teigia, kad tiesinės regresijos algoritmai parodė tikslesnius rezultatus užkodavus ciklinius duomenis, o sprendimų medžiai buvo neįtakoti. Verta plačiau išeksperimentuoti ar tai įtakoja kitus sudėtingesnius algoritmus.

2.2.1. Duomenų persidengimas

Norit tikslingai apmokinti dirbtino intelekto modelį svarbu suprasti kaip vartotojai elgiasi. Visa tai atsispindi duomenyse. Kaip matome iš (1 lentelė) nemaža dalis paieškų yra panašios tarp gretutinių mėnesių. Tai yra svarbu, kuriant naujas savybes iš kurių dirbtinio intelekto modeliai gali geriau apsimokinti.

1 lentelė. Įrašų persidengimas tarp gretutinių mėnesių pirmuose 10 000 įrašų.

Mėnesiai	**Abejuose mėnesiuose %**	**Tik antrame mėnesyje %**
2022-01 prieš 2022-02	59.510915	20.934063
2022-02 prieš 2022-03	43.230361	28.709146
2022-03 prieš 2022-04	51.899487	25.441232
2022-04 prieš 2022-05	46.626608	27.162836
2022-05 prieš 2022-06	36.658092	32.795447
2022-06 prieš 2022-07	29.228613	36.568686
2022-07 prieš 2022-08	27.592846	36.175766
2022-08 prieš 2022-09	34.957510	35.200922
2022-09 prieš 2022-10	47.342093	26.663614
2022-10 prieš 2022-11	52.616058	23.899670
2022-11 prieš 2022-12	33.031569	34.629372
2022-12 prieš 2023-01	45.756278	28.283356
2023-01 prieš 2023-02	56.012298	22.446481

2023-02 prieš 2023-03	40.093956	30.235597
2023-03 prieš 2023-04	34.305518	33.701959
2023-04 prieš 2023-05	29.821659	35.079083
2023-05 prieš 2023-06	12.476079	43.673279
2023-06 prieš 2023-07	17.691502	41.568431
2023-07 prieš 2023-08	31.576736	34.157001
2023-08 prieš 2023-09	24.924412	38.578502
2023-09 prieš 2023-10	31.186372	34.516325
2023-10 prieš 2023-11	34.156395	32.596484
2023-11 prieš 2023-12	42.101363	28.593657

Labai didelis persidengimas matomas vien tarp įrašų. Tęsiant analizę galima pasižiūrėti kiek oro uosto krypčių persidengia tarp gretutinių mėnesių.

2 lentelė. Oro uostų persidengimas tarp gretutinių mėnesių pirmuose 5000 įrašų.

Mėnesiai	**Oro uostai abejuose mėnesiuose %**	**Oro uostai tik antrame mėnesyje %**
2022-01 prieš 2022-02	98.205242	0.916127
2022-02 prieš 2022-03	97.869934	1.104618
2022-03 prieš 2022-04	96.894806	1.764630
2022-04 prieš 2022-05	96.315241	1.567154
2022-05 prieš 2022-06	97.093360	1.450983
2022-06 prieš 2022-07	96.512677	1.785034
2022-07 prieš 2022-08	97.151812	1.445555
2022-08 prieš 2022-09	96.817377	1.496287
2022-09 prieš 2022-10	97.177755	1.298357
2022-10 prieš 2022-11	97.449688	1.282132
2022-11 prieš 2022-12	94.289068	3.144350
2022-12 prieš 2023-01	94.937850	2.309966
2023-01 prieš 2023-02	98.831703	0.491110
2023-02 prieš 2023-03	98.436284	0.944562
2023-03 prieš 2023-04	97.435220	1.257312
2023-04 prieš 2023-05	95.534522	2.562642
2023-05 prieš 2023-06	95.298063	1.787754
2023-06 prieš 2023-07	97.906109	1.201172
2023-07 prieš 2023-08	97.775813	1.139572
2023-08 prieš 2023-09	96.709113	1.889640
2023-09 prieš 2023-10	96.961788	1.696940
2023-10 prieš 2023-11	97.384253	1.194364

2023-11 prieš 2023-12	98.090075	0.977974

Didelis persidengimas tarp gretutinių oro uostų (žr. 2 lentelė) reiškia prognozuojant į ateitį galima imti praeito mėnesio oro uostų vertes ir išgauti didelį verčių pasikartojimą.

2.3. Vertinimo metrikos

Laiko eilučių prognozės regresiniuose uždaviniuose dažniausiai naudojamos paklaidos: RMSE, MAE ir MAPE. Šios paklaidos skirtos parodyti nuokrypį nuo prognozuotų duomenų ir tikrų duomenų.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \bar{y_i}| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y_i})^2} \quad (2)$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \quad (3)$$

Vidurkinė kvadratinė paklaida (RMSE) apskaičiuojama kaip kvadratų sumos šaknis iš skirtumų tarp prognozuotos ir tikrosios vertės, padalinta iš prognozių skaičiaus. Tai dažnai naudojama, norint apibrėžti bendrą prognozuotos ir faktinės vertės skirtumą. Tuo tarpu, vidurkinė absoliuti paklaida (MAE) nustatoma kaip kiekvienos prognozuotos vertės absoliutaus skirtumo vidurkis nuo visų prognozių vidurkio. Vidurkinė absoliuti procentinė paklaida (MAPE) naudojama įvertinti modelio tikslumą prognozuojant reikšmes, atsižvelgiant į tikrosios vertės skaitinį dydį. Šios metrikos parodys modelių prognozavimo tikslumą.

3. Skrydžių prognozavimo eksperimentiniai tyrimai

3.1. Prognozė pagal išvykimo datą

Šiai eksperimentų skilčiai sudaroma supaprastinta situacija, siekiant pastebėti ar aprašomas sprendimas įgyvendinamas. Imama viena oro uostų kryptis ir jos išvykimo data.

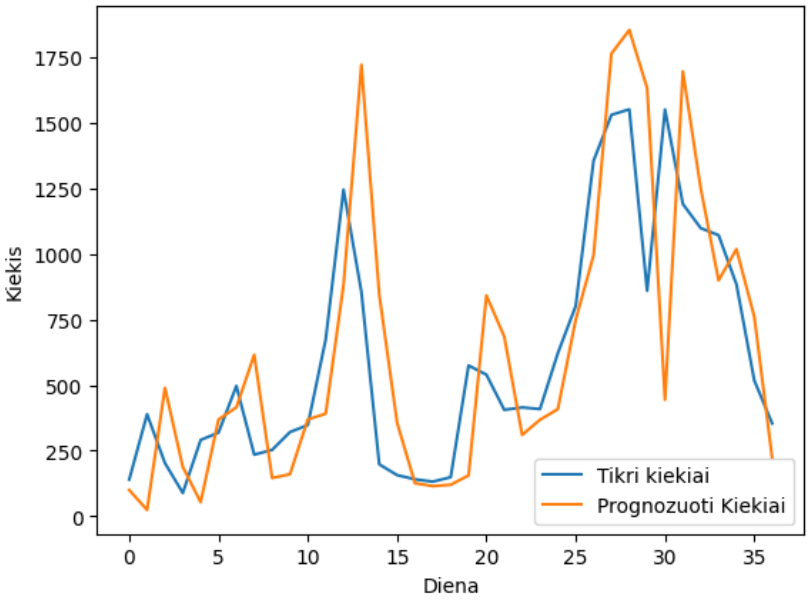
Kryptis	Visas duomenų rinkinys	Išvykimo datos (treniravimo)	Išvykimo datos (testavimo)
Niujorkas-Majamis	2021-01-01 – 2024-01-01	2023-01-01 – 2023-11-25	2023-11-25 – 2024-01-01

3.1.1. Statistiniai modeliai

Statistiniai modeliai yra vieni iš populiariausių pasirinkimų laiko eilučių prognozavimo uždaviniuose dėl jų greičio, paprastumo ir tikslumo. Jie dažnai vertinami dėl galimybės greitai apdoroti duomenis ir gauti tikslų rezultatą, kuris yra svarbus įvairiose pramonės srityse ir verslo sprendimuose.

ARIMA

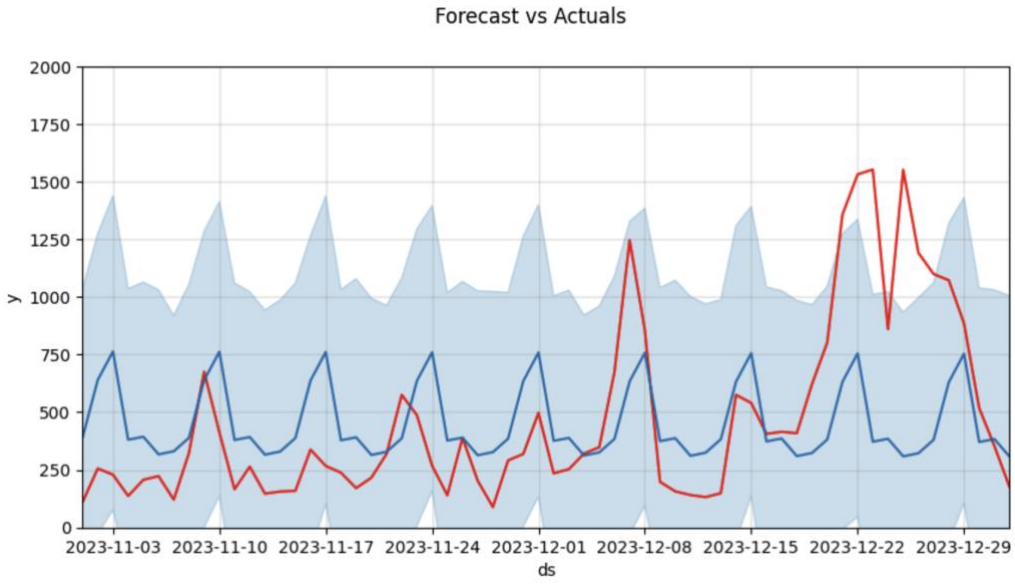
ARIMA modelio apmokinimui buvo naudojami duomenys iki 2023-11-25 dienos. Modelio testavimo rezultatai rodo, kad prognozuojamos vertės seka duomenų tendenciją, tačiau yra nutolusios nuo tikrų verčių (žr. 14 pav.).



14 pav. ARIMA modelio testavimo duomenų prognozių palyginimas su tikrais duomenimis.

Facebook „Prophet“

Facebook „Prophet“ (liet. „Pranašas“) modelis buvo specialiai sukurtas laiko eilučių uždaviniams. Šis modelis geba išgauti tendencijas, sezoniškumus, nuokrypius ir panašias savybes iš pateiktų duomenų. Aišku tokias savybes pagauti didelio kiekio duomenų. Kaip matome iš (žr. 15 pav.) nors modelis mokintas ant beveik trijų metų duomenų (2021 – 2023), modelis nustato bendrą duomenų tendenciją ir ją pateikia kaip prognozės atsakymą. Nors prognozė (mėlyna linija) kartais sutampa su tikrais duomenimis (raudona linija) to neužtenka užtikrinti tikslumo ir duomenų sutapimo.



15 pav. Prophet modelio prognozės palyginimas su tikrais duomenimis.

3.1.2. Neuroniniai tinklai

Palyginti su statistiniais metodais, neuroniniai tinklai gali panaudoti įvairias savybes, neapsiribodami vien tik istoriniais kiekiais ar regresiniu pasikartojimu. Neuroniniai tinklai gali išmokti asociacijas tarp paieškų duomenų, kurių standartiniai algoritmai neaptinka: delta tarp laiko intervalų, panašių oro uostų paieškų vertės, metinis paieškų pasikartojimas ir kitos svarbios savybės. Dėl šios priežasties neuroninių tinklų struktūroms paduodamos papildomos savybės (žr. lentelė).

3 lentelė. Pildomos savybės neuroninio tinklo treniravimui.

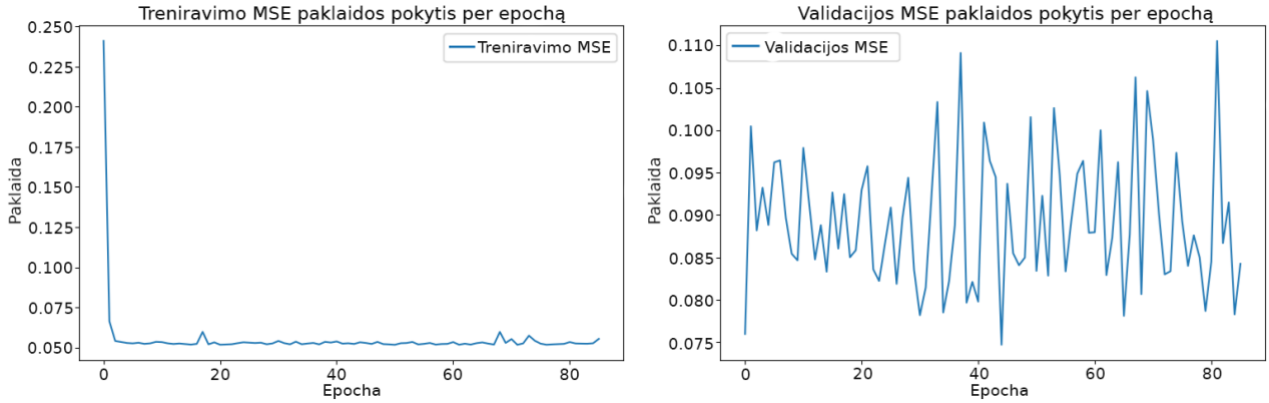
Savybė	**Aprašas**
yoy_cnt	Dieną prieš metus buvęs paieškų kiekis.
cnt-N	N – dienos. Prieš N dienų buvęs paieškų kiekis.
ot_cnt-N	N – dienos. Prieš N dienų buvęs paieškų kiekis priešingoje skrydžių kryptyje (Majamis-Niujorkas)

Tiesinis neuroninis tinklas

Atliekamas eksperimentas su tiesiniais neuroniniais tinklais, kurių struktūra matoma (žr. 16 pav.). Tai neuroninis tinklas sujungtas tiesiniais sluoksniais, tarp kurių yra ReLU aktyvacijos funkcija.

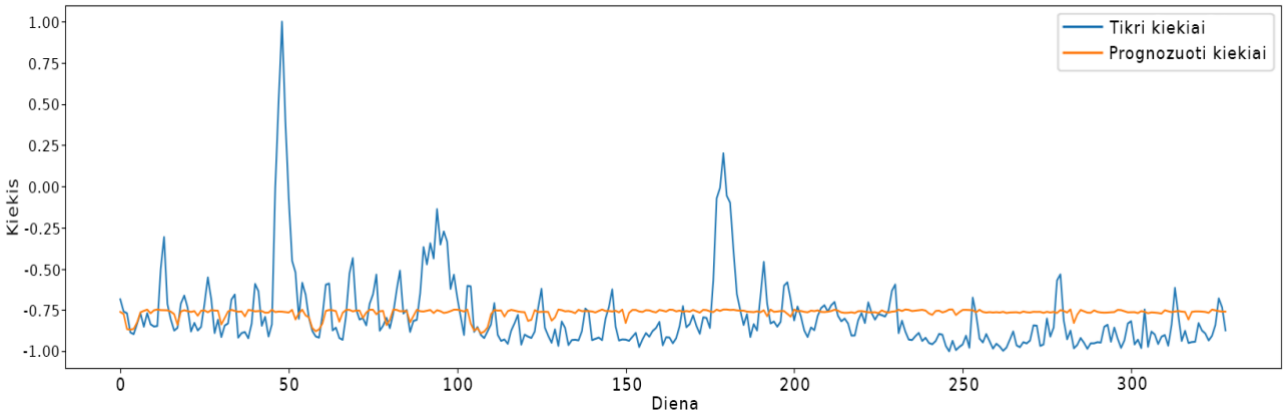
```
SimpleNN(
  (fc1): Linear(in_features=1, out_features=64, bias=True)
  (relu1): ReLU()
  (fc2): Linear(in_features=64, out_features=32, bias=True)
  (relu2): ReLU()
  (fc3): Linear(in_features=32, out_features=32, bias=True)
  (relu3): ReLU()
  (fc4): Linear(in_features=32, out_features=1, bias=True)
)
```

16 pav. Tiesinio neuroninio tinklo struktūra.

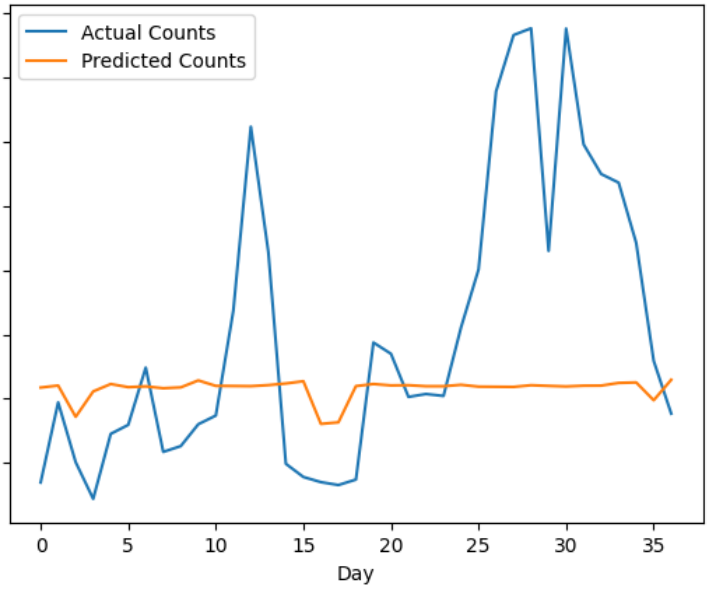


17 pav. Tiesinio neuroninio tinklo treniravimo ir validacijos paklaidos pokytis epochų atžvilgiu.

Iš (žr. 17 pav.) matome, kad modelis nesimokina pateiktų savybių, paklaida nekinta ir modelio prognozės tikslumas nedidėja. Tai ypač matoma validacijos paklaidos grafike, nes paklaida stipriai varijuoja, bet išlieka aukšta. Galima matyti (žr. 18 pav.), kad modelis nesugeba tiksliai prognozuoti treniravimo duomenų.



18 pav. Neuroninio tinklo treniravimo duomenų prognozių palyginimas su tikrais duomenimis.

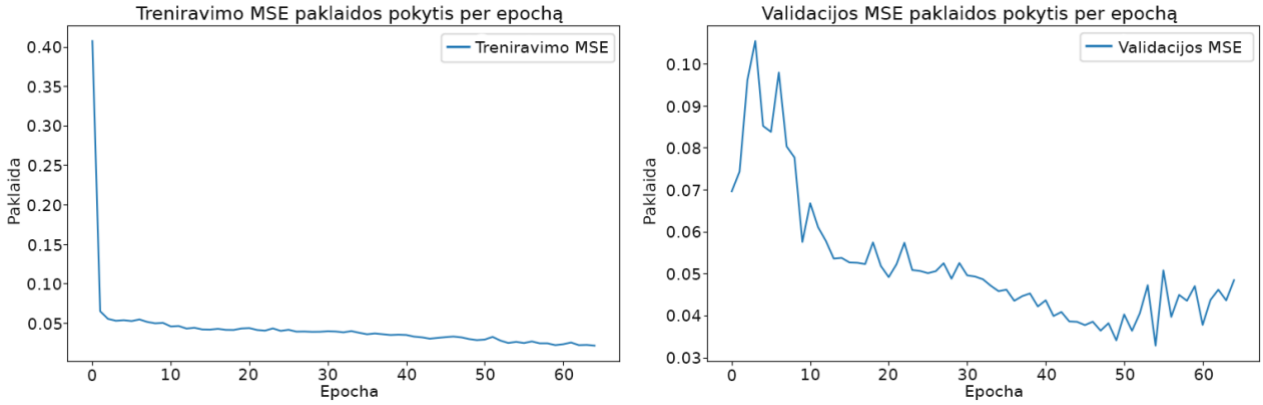


19 pav. Neuroninio tinklo testavimo duomenų prognozių palyginimas su tikrais duomenimis.

Validacijos ir tikslumo rodikliai rodo, kad ši neuroninio tinklo architektūra neapsimokina ir negali tiksliai prognozuoti skrydžių paieškų.

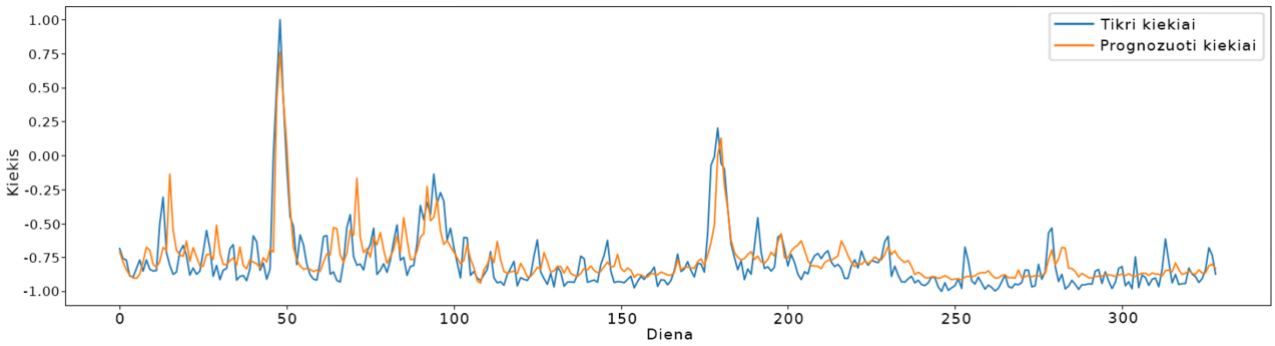
LSTM

LSTM modelio atminties valdymo mechanizmai leidžia tinklui išlaikyti informaciją ilgiau, tai yra svarbu modeliuojant pasikartojančius vartotojų elgsenos bruožus.

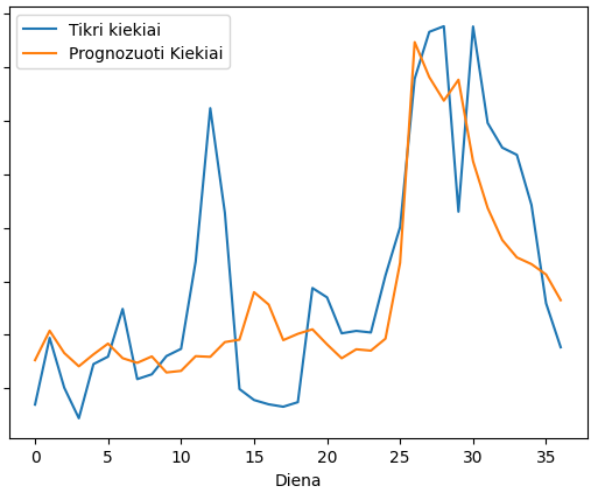


20 pav. LSTM modelio treniravimo ir validacijos paklaidos pokytis epochų atžvilgiu.

Matome iš (žr. 20 pav.), kad modelis mokosi iš duomenų, mažėja paklaida tiek treniravimo, tiek validacijos duomenyse. Po apmokymo LSTM modelio prognozė (žr. 21 pav.) glaustai seka tikras duomenų vertes.



21 pav. LSTM tinklo treniravimo duomenų prognozių palyginimas su tikrais duomenimis.

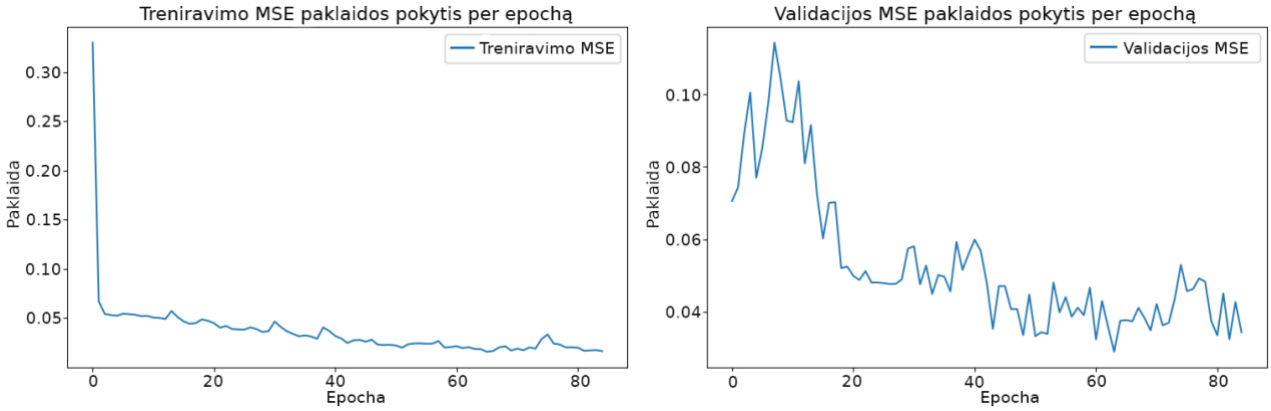


22 pav. LSTM testavimo duomenų prognozių palyginimas su tikrais duomenimis.

Testavimo prognozėje LSTM modelio prognozuoti kiekiai gana stipriai skiriasi nuo tikrųjų reikšmių, bet yra arti bendros duomenų tendencijos.

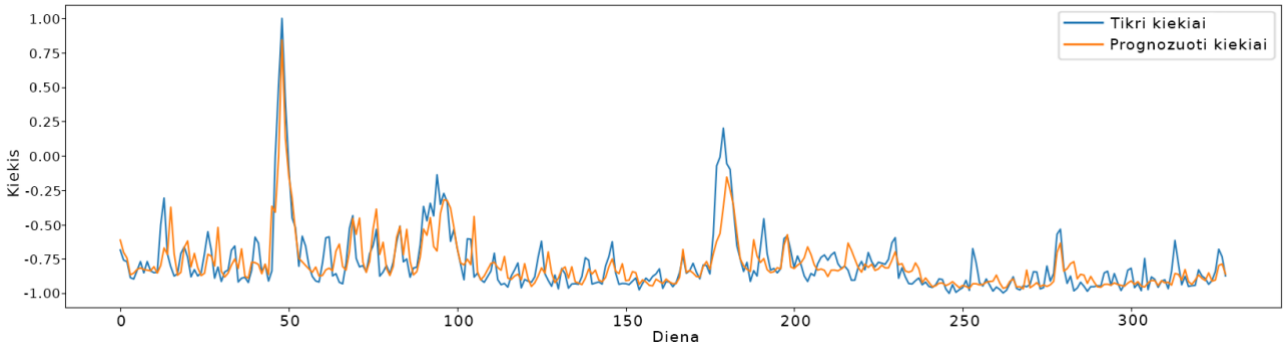
Seq2Seq

Seq2Seq yra vienas naujesnių modelių su užkodavimo ir atkodavimo blokais. Šiuo atvejų užkodavimo ir atkodavimo blokai su daryti iš LSTM modulių.

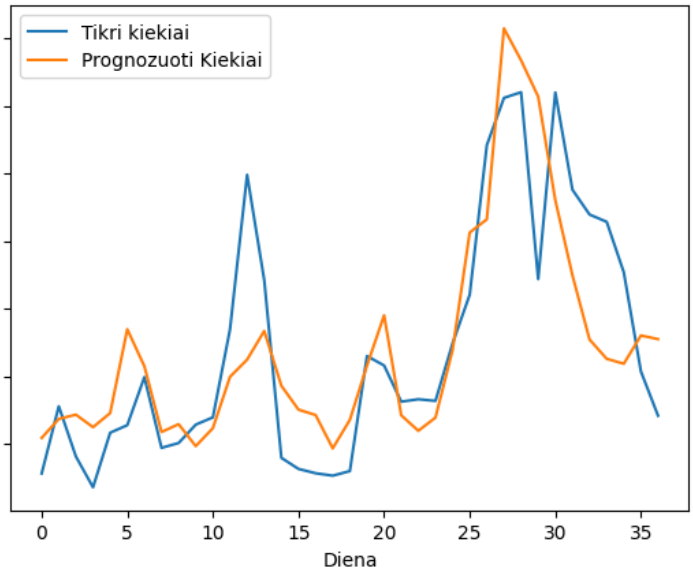


23 pav. Seq2Seq modelio treniravimo ir validacijos paklaidos pokytis epochų atžvilgiu.

Grafikas (žr. 23 pav.) rodo, kad modelis mokinasi savybes, kadangi po truputį mažėja tiek treniravimo, tiek validacijos paklaida. Apmokinto modelio prognozės yra nenutolusios nuo treniravimo duomenų tendencijos ir apgaubia, net staigius duomenų pokyčius (žr. 24 pav.).



24 pav. Seq2Seq tinklo treniravimo duomenų prognozių palyginimas su tikrais duomenimis.



25 pav. Seq2Seq testavimo duomenų prognozės palyginimas su tikrais duomenimis.

4 lentelė. Išbandytų modelių paieškų prognozės vertinimo rodikliai iš visų testavimo duomenų prognozių.

	MAE	**RMSE**	**RMSLE**	**MAPE**
Seq2Seq	**210.4217**	**267.8561**	**0.5065**	**52.3841**
LSTM	225.3396	286.7298	0.5994	59.7343
DNT	341.5697	462.6684	0.7611	74.0192
ARIMA	265.6665	362.1653	0.7814	56.7277

Pažvelgus į testavimo rezultatus matome dar mažesnę paklaidą negu LSTM modelis. Tai rodo Seq2Seq modelio pranašumą šiame uždavinyje.

3.2. Skrydžių prognozė pagal paieškos datą

Ankstesnio skyriaus rezultatai parodė, kad pasirinkus išvykimo datą kaip laiko dimensiją, statistiniai metodai ir dirbtinio neuroninio tinklo modeliai prognozuoja paieškų kiekius, tačiau kelionių įrankio vartotojai, kaip ir daugelis žmonių planuoja keliones kelias dienas ar net mėnesius į priekį. Įtraukus kelionės paieškos datą kaip pagrindinę laiko vertikalę galima prognozuoti kaip kelionės paieškos kiekiai keičiasi artėjant iki išvykimo datos. Paieškos datos įtraukimas sukuria įdomią laiko eilučių variaciją, kur kiekviena kelionės kryptis, išvykimo ir atvykimo data, turi atskirą laiko eilučių duomenų rinkinį. Tokia variacija reikalauja modeliams paduoti daug savybių (žr. 5 lentelė), o daugelis statistinių modelių yra nepajėgus apdoroti daugiau negu vieno parametro ir gali prognozuoti vertes tik į artimą ateitį, iškart po žinomų duomenų pabaigos. Dėl šių priežasčių tyrimams bus naudojami dirbtinių neuroninių tinklų modeliai.

5 lentelė. Savybės naudojamos treniruoti modelius skrydžių prognozėms pagal paieškos datą.

Savybė	**Aprašas**
Paieškų kiekis	Kiekis ieškotos oro uostų ir datų kombinacijos
Išvykimo oro uosto platuma	Išvykimo oro uosto koordinatės
Išvykimo oro uosto ilguma	
Atvykimo oro uosto platuma	Atvykimo oro uosto koordinatės
Atvykimo oro uosto ilguma	
Delta iki išvykimo datos	Dienų skirtumas nuo paieškos datos iki išvykimo datos
Delta iki atvykimo datos	Dienų skirtumas nuo paieškos datos iki atvykimo datos
Delta tarp išvykimo ir atvykimo datų	Dienų skirtumas nuo išvykimo datos iki atvykimo datos
Išvykimo diena metuose	Numeris atspindintis išvykimo datos dieną metuose. Sausio 1d. vertė būtų lygi 1
Atvykimo diena metuose	Numeris atspindintis atvykimo datos dieną metuose. Sausio 1d. vertė būtų lygi 1
Paieškos diena metuose	Numeris atspindintis paieškos datos dieną metuose. Sausio 1d. vertė būtų lygi 1

Šioje tyrimų skiltyje duomenys bus skaidomi į treniravimo ir testavimo rinkinius pagal paieškos data (žr. 6 lentelė.). Modeliai bus mokinami prognozuoti ateities vertes keletui skirtingų oro uostų, dėl to tarp pateiktų savybių (žr. 5 lentelė) matomos išvykimo ir atvykimo oro uostų koordinatės. Kiekvienas modelis gauna po 6 paskutinius laiko žingsnius, pagal kuriuos bus prognozuojami ateities paieškų kiekiai.

6 lentelė. Duomenų skaidymas pagal paieškos datą.

Visas duomenų rinkinys	Treniravimo rėžiai	Testavimo rėžiai
2021-01-01 – 2023-01-01	2021-01-01 – 2022-01-01	2022-01-01 – 2022-06-01

Literatūros analizė parodė, kad įvairūs dirbtinio neuroninio tinklo modeliai su aukštu tikslumu prognozuoja ateities vertes. Tarp aptartų straipsnių išsiskiria keli modeliai: paprastas neuroninis tinklas, LSTM ir Seq2Seq. Žemiau pateikiamos modelių variacijos bus naudojamos šios skilties bandymuose.

PDNT – neuroninis tinklas su 3 paslėptais sluoksniais, pirmas sluoksnis sudarytas iš 16 neuronų, antras ir trečias iš 32 neuronų. Tarp paslėptų sluoksnių parinkta ReLU aktyvacijos funkcija.

LSTM – neuroninis tinklas su vienu LSTM bloku, kuris turi 16 neuronų ir vienu paslėptu sluoksniu iš 32 neuronų. Tarp LSTM bloko ir paslėpto sluoksnio pasirinkta ReLU aktyvacijos funkcija.

Seq2Seq – neuroninis tinklas iš dviejų blokų: užkodavimo ir atkodavimo. Užkodavimo blokas sudarytas iš LSTM sluoksnio su 32 neuronais. Atkodavimo blokas sudarytas iš LSTM sluoksnio ir vieno paslėpto sluoksnio, abu turi po 32 neuronus.

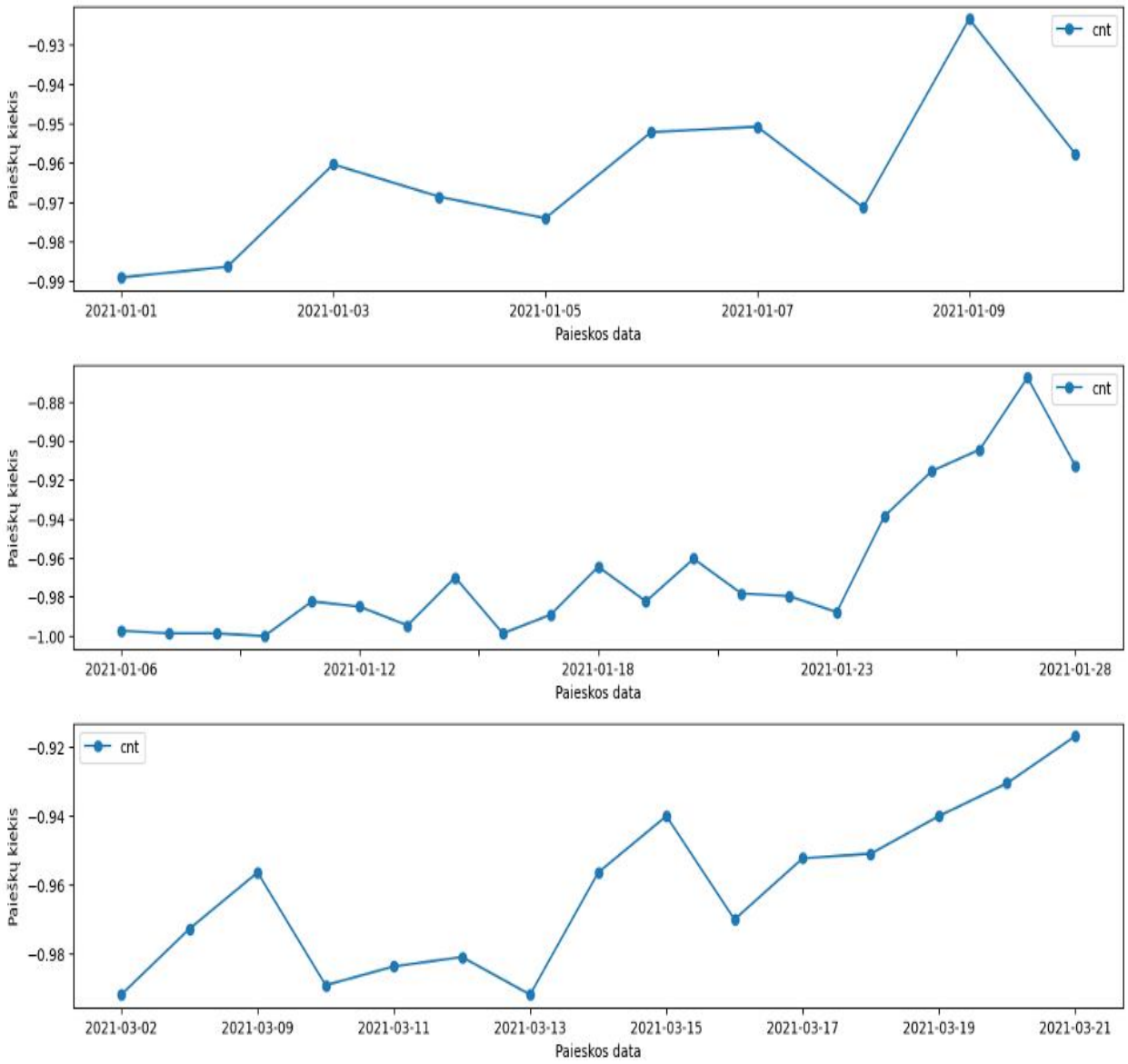
Eksperimentų metu išbandytos įvairios modelių treniravimo sąlygos: skirtingas epochų kiekis, mokymosi sparta, imties dydis. Rastos palankiausios sąlygos modelių mokymui. (žr. 7 lentelė). Modeliams pateikiamų duomenų mastelis sumažintas rėžiuose (-1, 1), siekiant sutrumpinti mokymosi ir skaičiavimų laiką.

7 lentelė. Modelių treniravimo sąlygos.

Epochų kiekis	60
Paklaidos funkcija	MSE
Ankstyvas stabdymas	Per 10 epochų nepagerinus paklaidos rezultato
Optimizavimo algoritmas	Adam
Imties dydis	8
Mokymosi sparta	0,001
Mokymosi spartos valdiklis	Kas 5 epochą mažinama mokymosi sparta 50%

3.2.1. Skrydžiai į vieną pusę

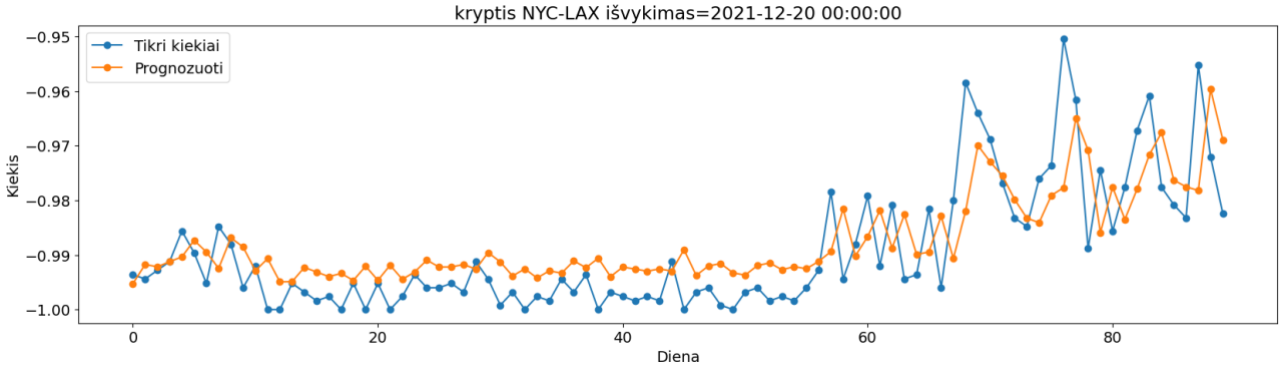
Didelė dalis vartotojų ieško skrydžių tik į vieną pusę arba skaido skrydžius į abi puses siekiant gauti geresnius pasiūlymus ar kainas. Šie vartotojų duomenys turi unikalius ryšius, dėl šios priežasties yra apmokinamas atskiras modelis. Skirtingai negu standartinis laiko eilučių uždavinys, kur dažniausiai yra viena ilga laiko vertikalė, kelionių paieškos turi daug mažų duomenų rinkinių kiekvienam oro uostui ir išvykimo datai (žr. 26 pav. Skirtingi laiko eilučių pavyzdžiai pagal išvykimo datą.). Dienų tarpai ir pradžios taškai gali skirtis nuo oro uosto derinio ir išvykimo datos.



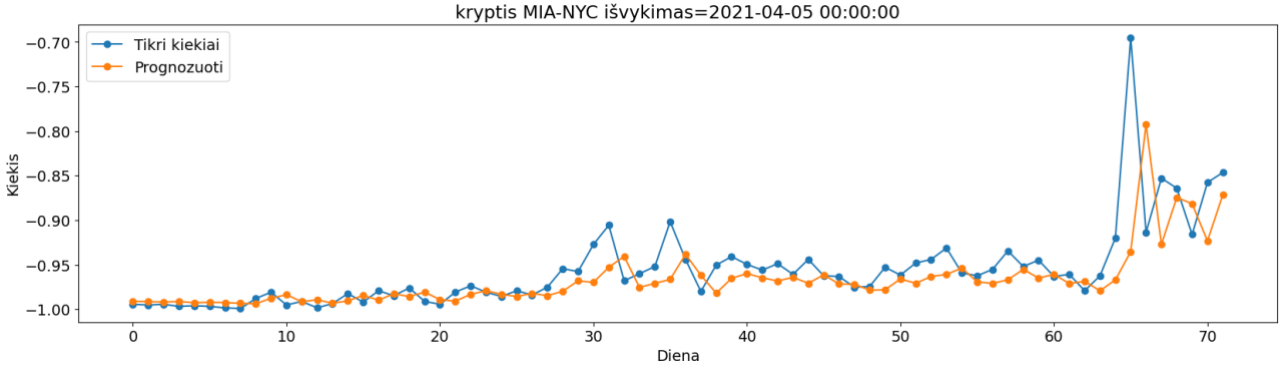
26 pav. Skirtingi laiko eilučių pavyzdžiai pagal išvykimo datą. Kryptis Niujorkas-Majamis.

3.2.2. Skrydžiai į vieną pusę – PDNT

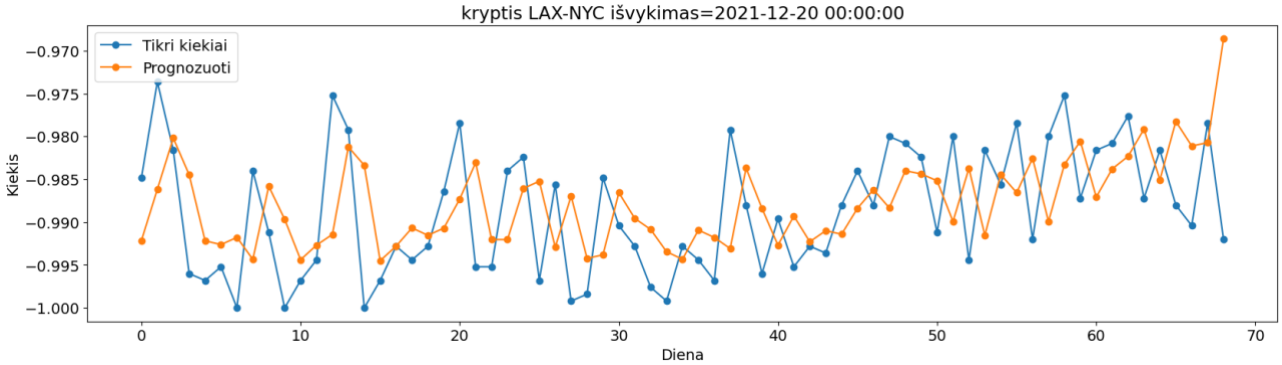
Iš pradžių bandymuose naudojamas PDNT modelis, kadangi jo struktūra yra gana paprasta ir nereikalauja intensyvaus kompiuterio darbo. Pradedant nuo paprasčiausio modelio, galima lengvai įvertinti, ar tikslioms prognozėms reikia sudėtingesnio algoritmo. Pirmiausia modelis buvo apmokytas prognozuoti vieną laiko žingsnį į priekį, tai pagal duomenų aprašą atitinka vieną dieną į ateitį. Analizuojant treniravimo duomenų pavyzdžius (žr. 27 pav.), pastebėta, kad kiekvienas oro uostų derinys turi skirtingas paieškų variacijas. Nors diagramoje paieškų kiekiai yra transformuoti į mažesnį mastelį, galima pastebėti, kad prognozuojant vieną laiko žingsnį į priekį, apmokytas PDNT modelio prognozės rezultatai yra arti tikrų verčių.



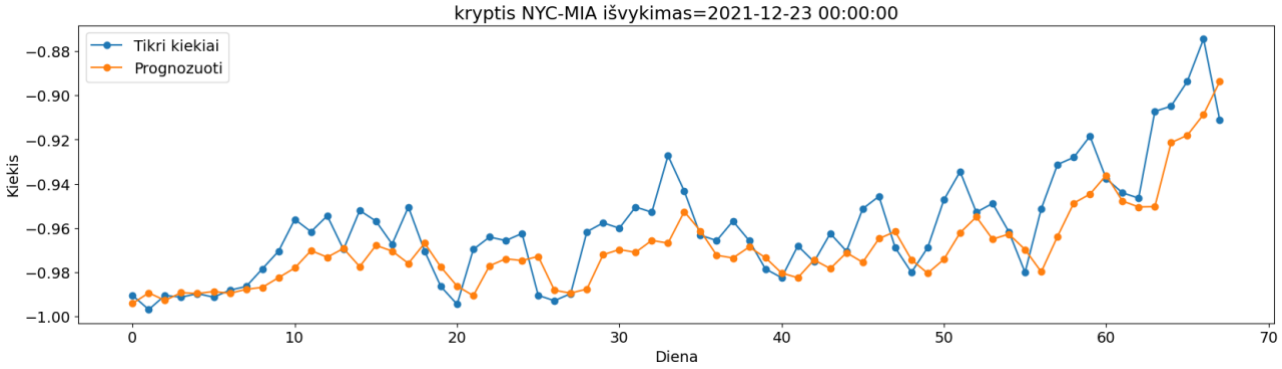
a) Kryptis Niujorkas – Los Andželas, paieškos artėjant iki išvykimo datos.



b) Kryptis Majamis – Niujorkas, paieškos artėjant iki išvykimo datos.



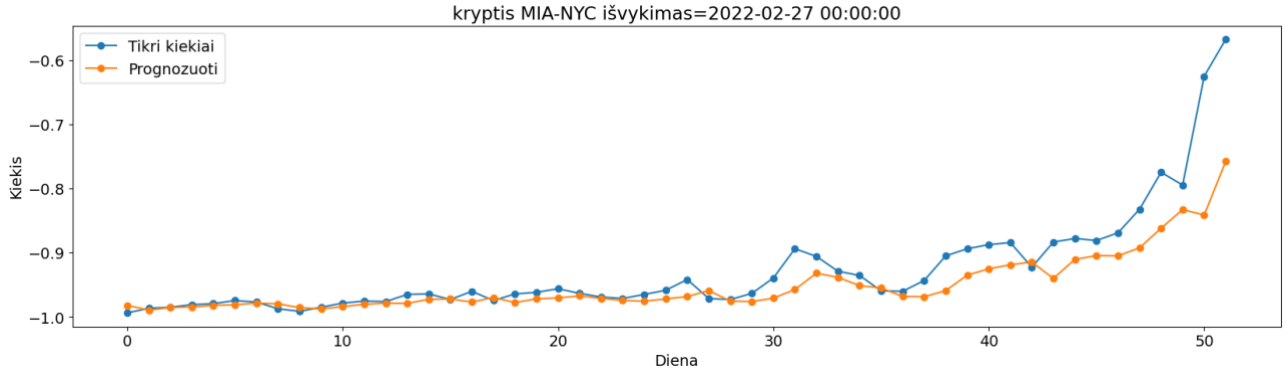
c) Kryptis Los Andželas – Niujorkas, paieškos artėjant iki išvykimo datos.



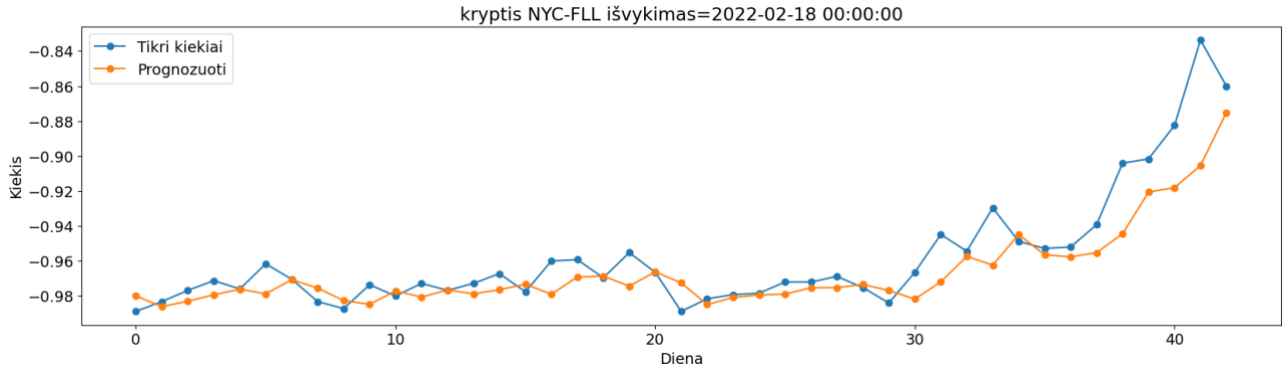
d) Kryptis Niujorkas – Majamis, paieškos artėjant iki išvykimo datos.

27 pav. PDNT kelių oro uostų sekančios dienos prognozės. Treniravimo duomenys. Tikri kiekai (Mėlyna) ir prognozuoti (oranžiniai).

Po PDNT modelio apmokymo svarbu patikrinti, kaip modelis prognozuoja pagal jam nežinomus testavimo duomenis. Testavimo duomenys parodo, kaip modelis elgsis realiose situacijose. Iš grafikų (žr. 28 pav.), matyti, kad modelio prognozės nedaug nutolę nuo tikrų duomenų.



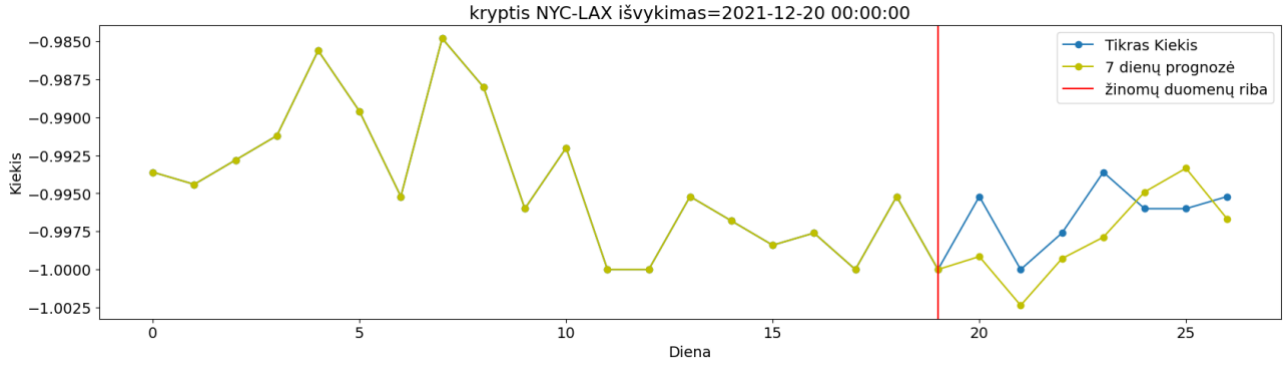
a) Kryptis Majamis – Niujorkas, paieškos artėjant iki išvykimo datos.



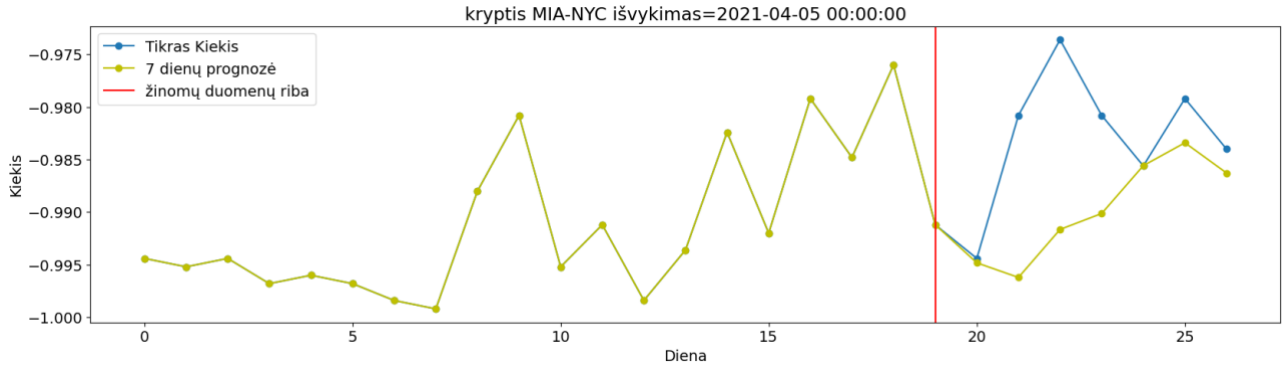
b) Kryptis Niujorkas – Fort Lauderdeilas, paieškos artėjant iki išvykimo datos.

28 pav. PDNT kelių oro uostų sekančios dienos prognozės. Testavimo duomenys. Tikri kiekai (Mėlyna) ir prognozuoti (oranžiniai).

Sekančios dienos prognozė, nesuteikia pakankamai laiko prognozuotus rezultatus tinkamai paruošti vartotojams. Kelių laiko žingsnių prognozė leistų apsvarstyti galimus scenarijus, izoliuoti dienas, kai paruošti rezultatai pasiektų didžiausią vartotojų kiekį. Modelis, mokytas prognozuoti tik sekantį laiko žingsnį, negali tiksliai prognozuoti kelių žingsnių į priekį. Siekiant prognozuoti skirtingus laiko žingsnius į ateitį, turi būti apmokintas atskiras modelis. Treniravimo duomenų prognozės pavyzdžiai (žr. 29 pav.) rodo, kad ši užduotis yra sudėtingesnė, nes per 7 dienų laiko tarpą duomenų tendencija gali žymiai pakisti. PDNT modelio testavimo grafikuose matoma, kad modelio prognozės rezultatai dažnai neatitinka tikrų duomenų (žr. 30 pav.).

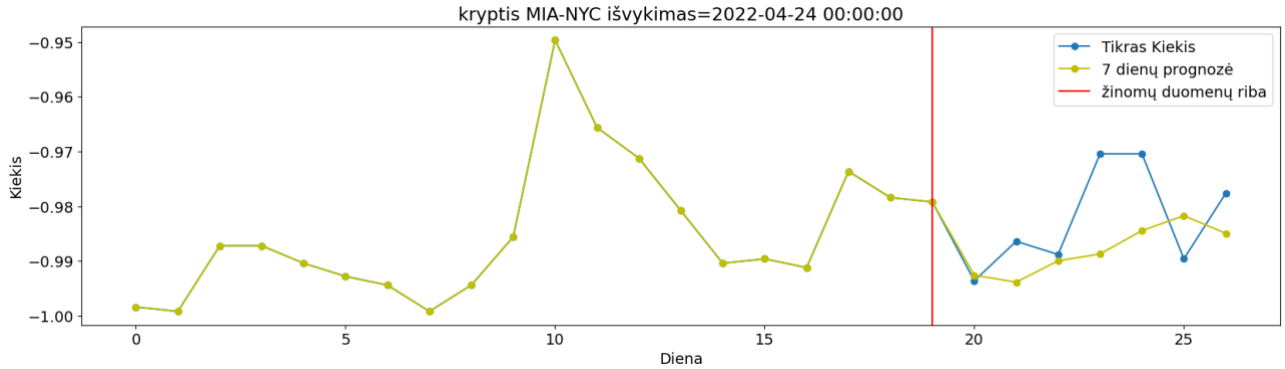


a) Kryptis Niujorkas – Los Andželas.

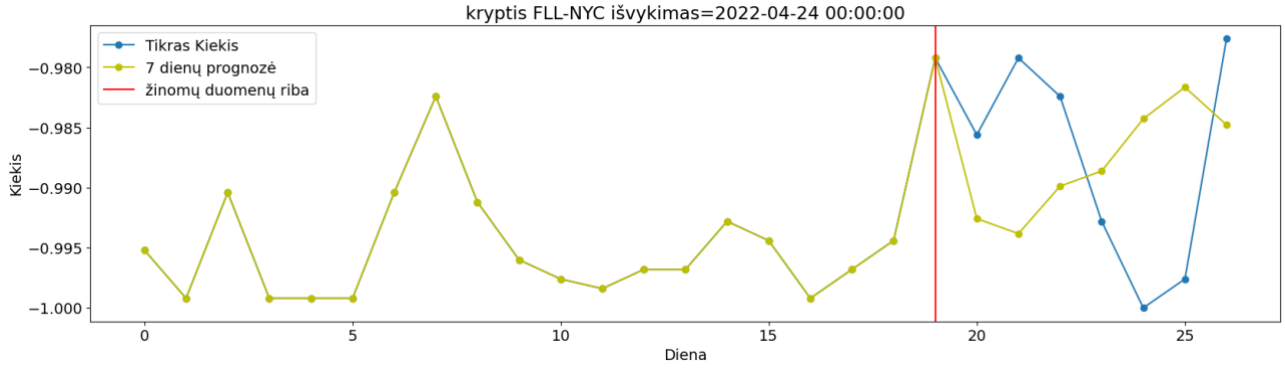


b) Kryptis Majamis – Niujorkas.

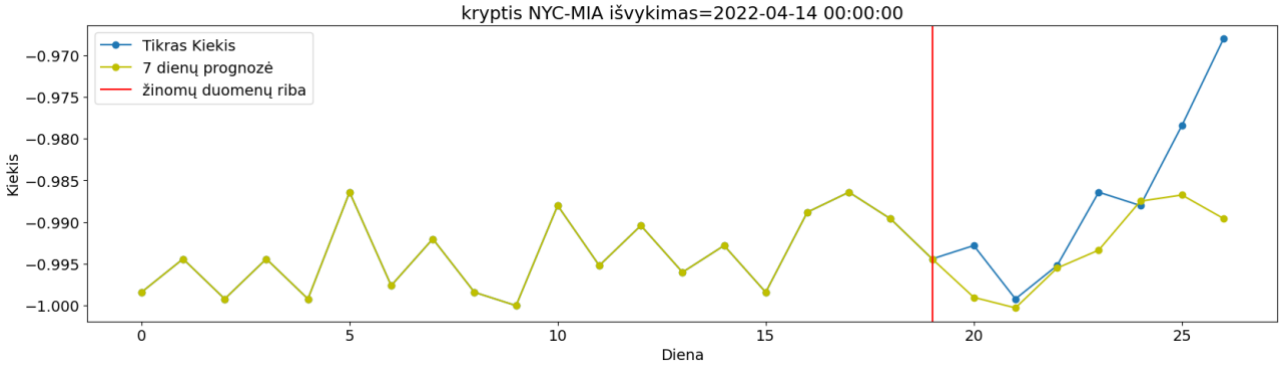
29 pav. PDNT kelių oro uostų sekančių 7 dienų prognozė nuo paskutinių žinomų duomenų taško. Treniravimo duomenys.



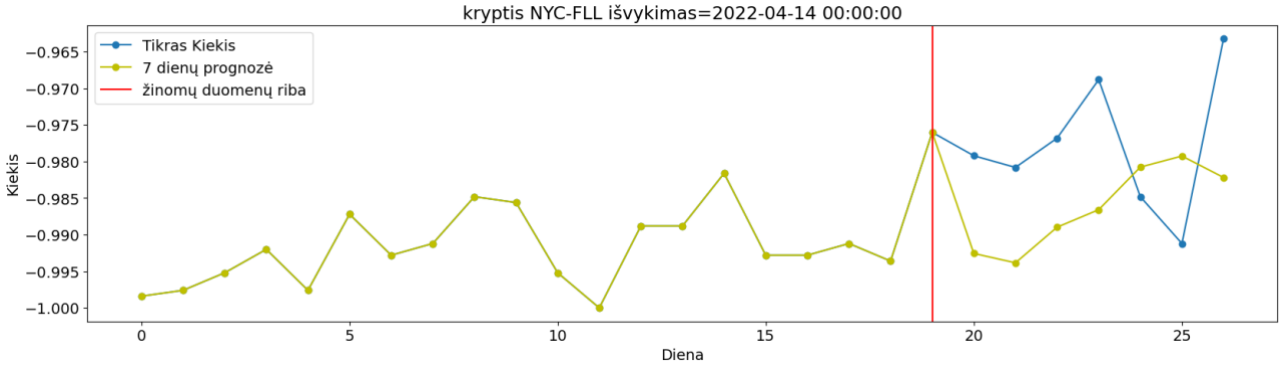
a) Kryptis Majamis – Niujorkas.



b) Kryptis Fort Lauderdeilas – Niujorkas.



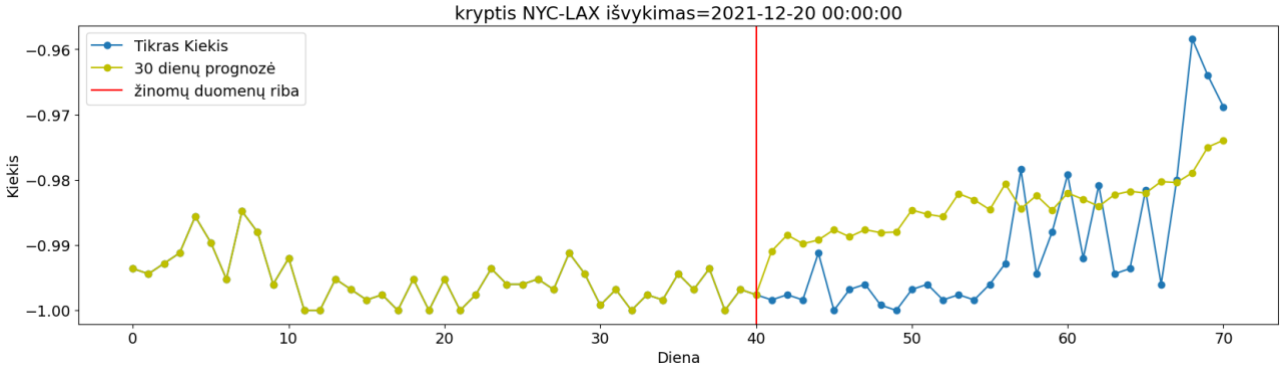
c) Kryptis Niujorkas – Majamis.



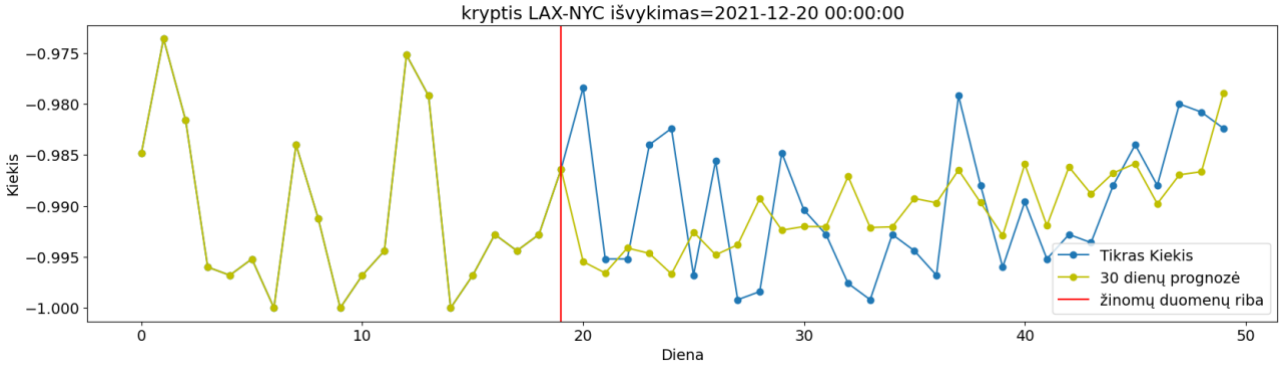
d) Kryptis Niujorkas – Fort Lauderdeilas.

30 pav. PDNT kelių oro uostų sekančių 7 dienų prognozė nuo paskutinių žinomų duomenų taško. Testavimo duomenys.

Kelių dienų prognozė gali stipriai įtakoti strateginius planus, tačiau rimtai paruošti kelias strategijas reikalinga 30 dienų prognozė. Prognozuoti 30 dienų į priekį yra sudėtingas uždavinys, nes per tokį laiko tarpą galimas didelis duomenų pokytis. PDNT modelis buvo mokomas prognozuoti 30 laiko žingsnių į priekį. Treniravimo prognozių pavydžiai (žr. 31 pav.) rodo, kad modelis netiksliai prognozuoja paieškų kiekius 30 dienų į priekį, tačiau nenukrypsta per daug nuo duomenų tendencijos, tai patvirtina testavimo prognozių grafikai (žr. 32 pav.). Netikslūs rezultatai gali būti dėl kelių priežasčių: per mažos treniravimo duomenų imties, pernelyg paprastos modelio struktūros ar didelės duomenų įvairovės.

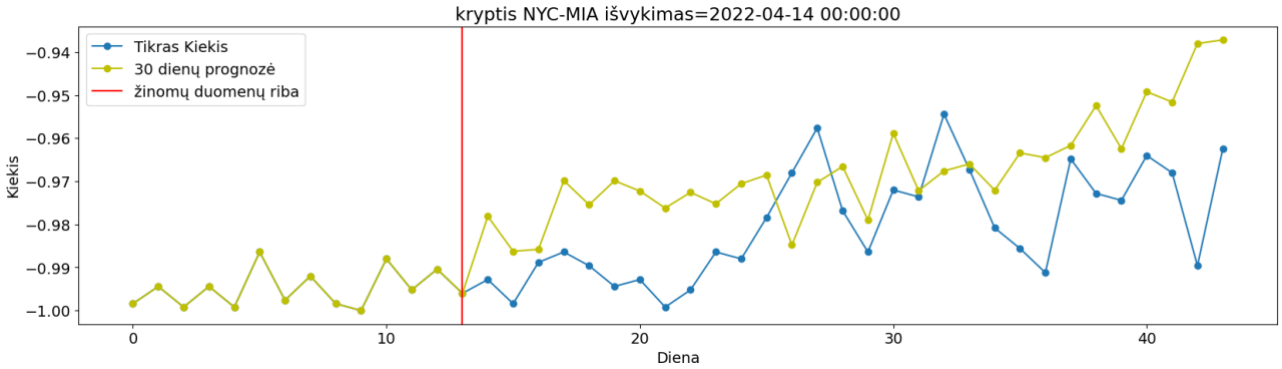


a) Kryptis Niujorkas – Los Andželas.

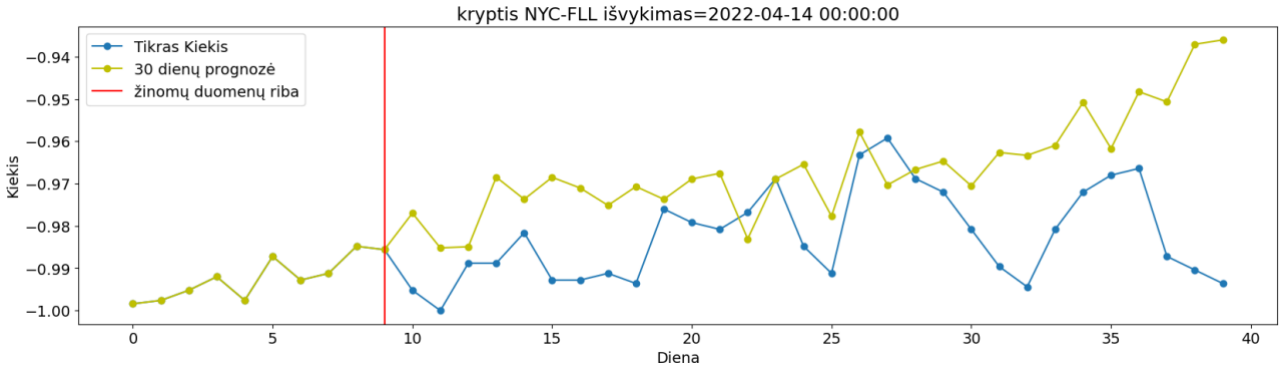


b) Kryptis Los Angelas – Niujorkas.

31 pav. PDNT kelių oro uostų sekančių 30 dienų prognozė nuo paskutinių žinomų duomenų taško. Treniravimo duomenys.



a) Kryptis Niujorkas – Majamis.



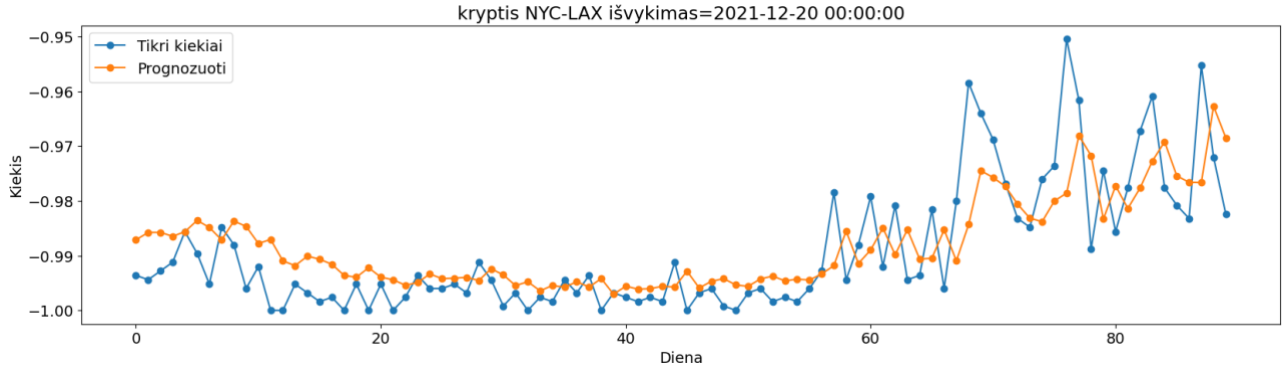
b) Kryptis Niujorkas – Fort Lauderdeilas.

32 pav. PDNT kelių oro uostų sekančių 30 dienų prognozė nuo paskutinių žinomų duomenų taško. Testavimo duomenys.

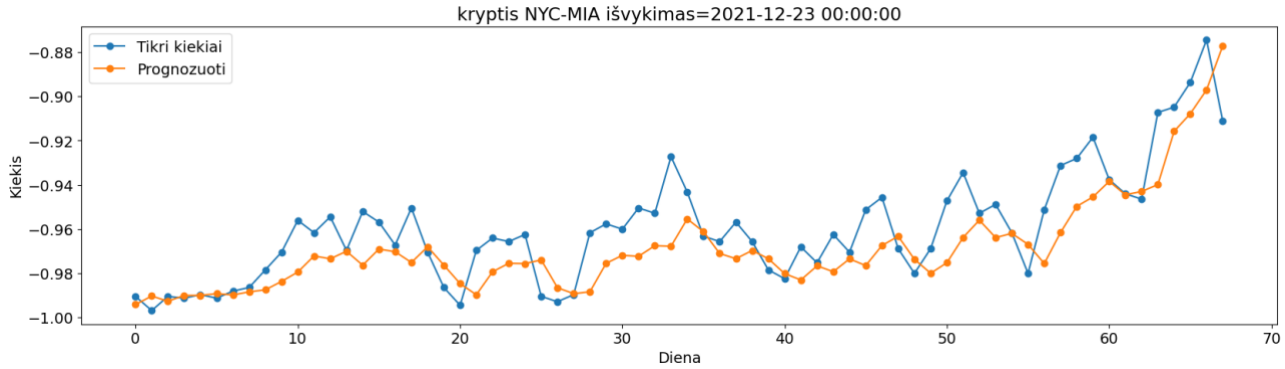
3.2.3. Skrydžiai į vieną pusę – LSTM ir Seq2Seq

Pirminių bandymų rezultatai parodė, kaip apmokintas PDNT modelis prognozuoja vienpusių skrydžių paieškas 1,7 ir 30 laiko žingsnių į priekį. Siekiant gauti tikslesnes prognozes, atliekami tyrimai su LSTM ir Seq2Seq modeliais, kurie yra paremti rekurentiniais neuroniniais tinklais. Sudėtingesnės struktūros modeliams dažnai reikia daugiau duomenų tinkamai sukalibruoti svorius. Iš LSTM modelio treniravimo prognozės grafikų (žr. 33 pav.) galima matyti, kad modelio prognozė

yra netoli tikrųjų paieškos kiekių. LSTM Modelis taip pat gerai prognozuoja paieškas testavimo duomenų grafikuose (žr. 34 pav.).

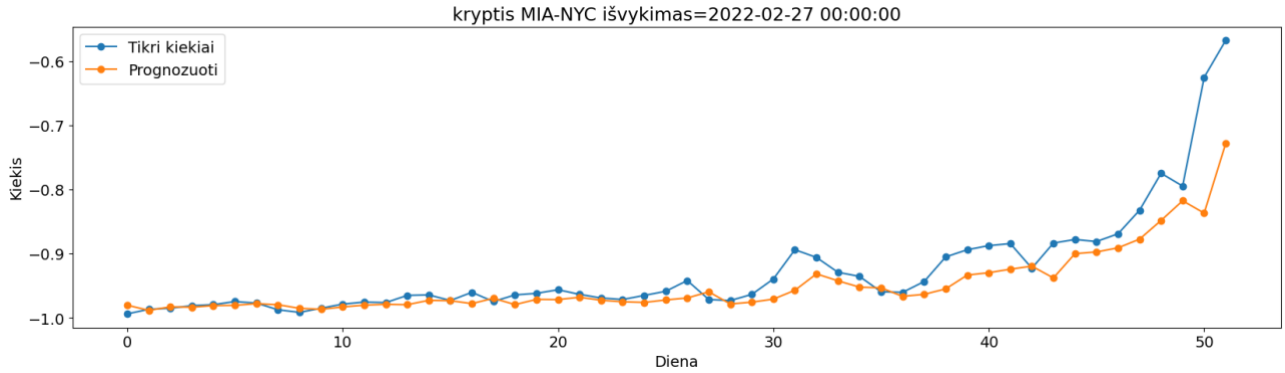


a) Kryptis Niujorkas – Los Andželas, paieškos artėjant iki išvykimo datos.

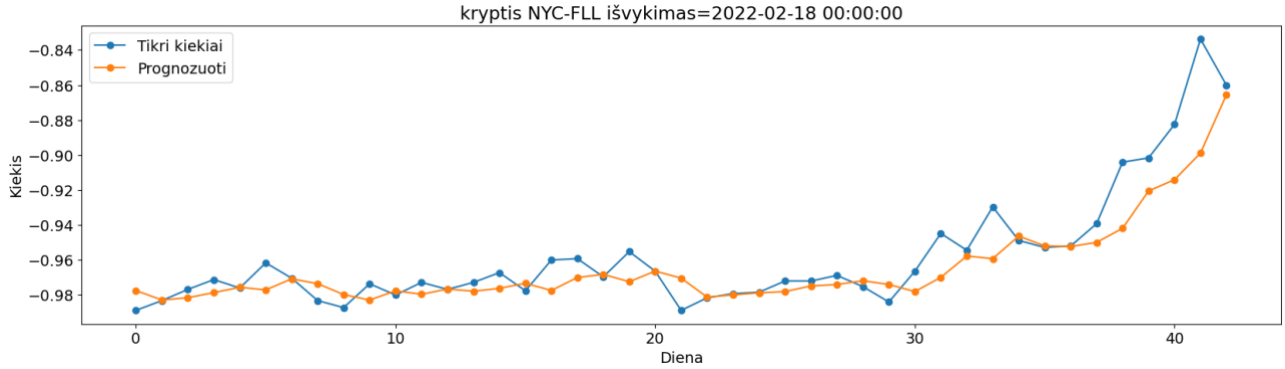


b) Kryptis Niujorkas – Majamis, paieškos artėjant iki išvykimo datos.

33 pav. LSTM kelių oro uostų sekančios dienos prognozės. Treniravimo duomenys.



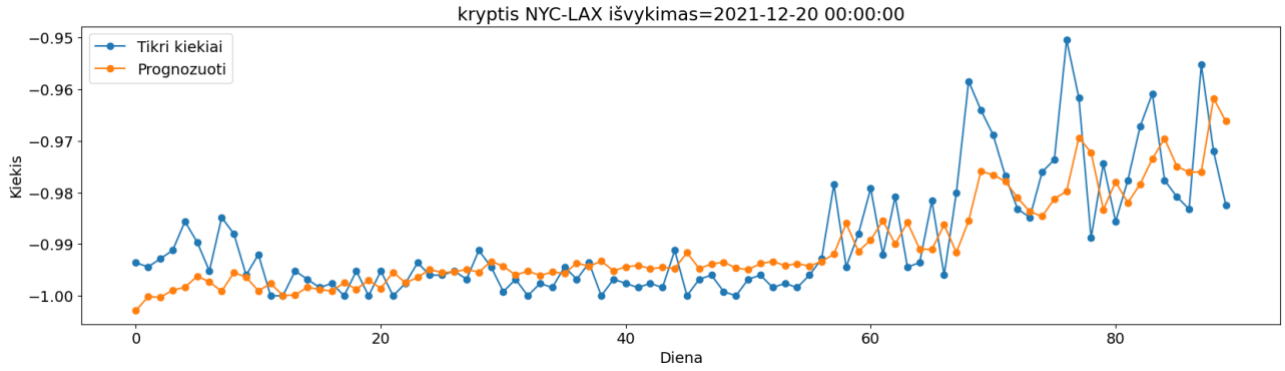
a) Kryptis Majamis – Niujorkas, paieškos artėjant iki išvykimo datos.



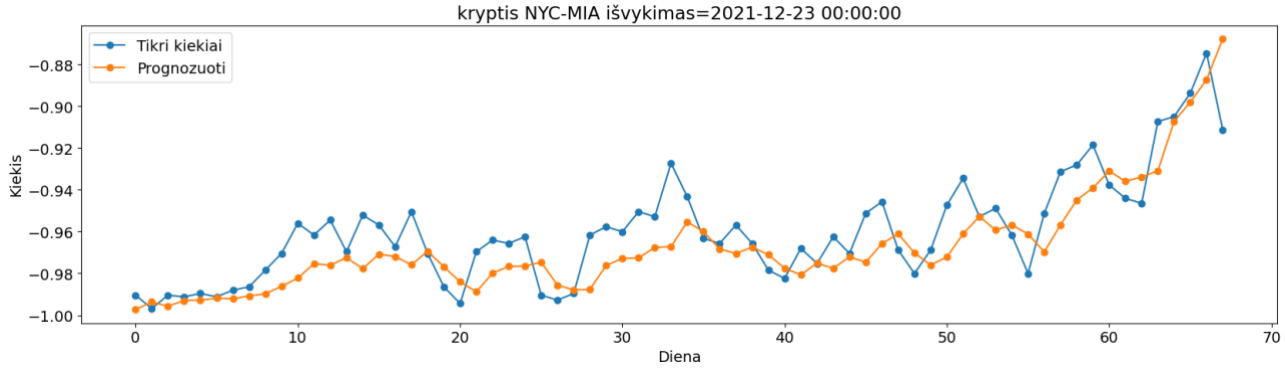
b) Kryptis Niujorkas – Fort Loderdeilas, paieškos artėjant iki išvykimo datos.

34 pav. LSTM kelių oro uostų sekančios dienos prognozės. Testavimo duomenys. Tikri kiekai (Mėlyna) ir prognozuoti (oranžiniai).

Nors Seq2Seq yra naujesnės architektūros modelis iš treniravimo ir testavimo prognozių (žr. 35 pav., 36 pav.) nėra aišku ar Seq2Seq duoda geresnį rezultatą negu LSTM. Nors tinkamesnio modelio išrinkti neišeina, galima pastebėti, kad abiejų modelių rezultatai yra arti tikrų paieškos verčių.

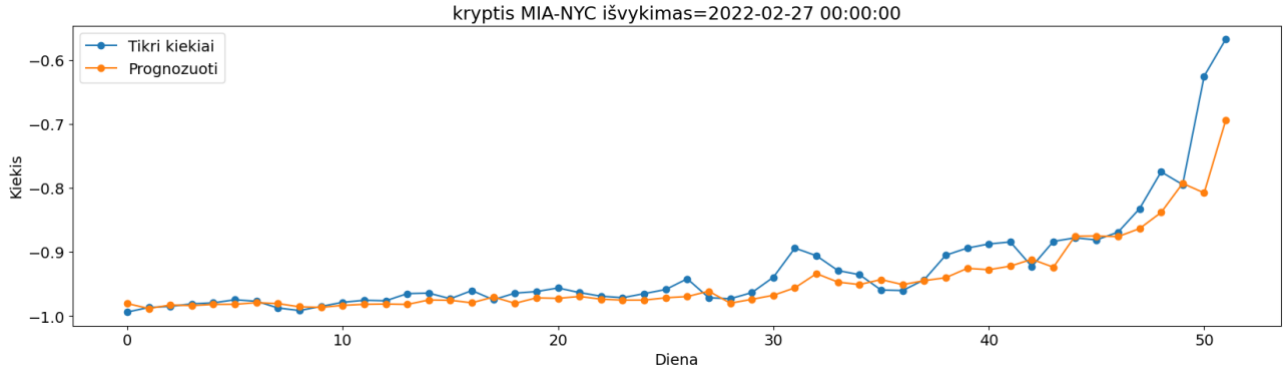


a) Kryptis Niujorkas – Los Andželas, paieškos artėjant iki išvykimo datos.

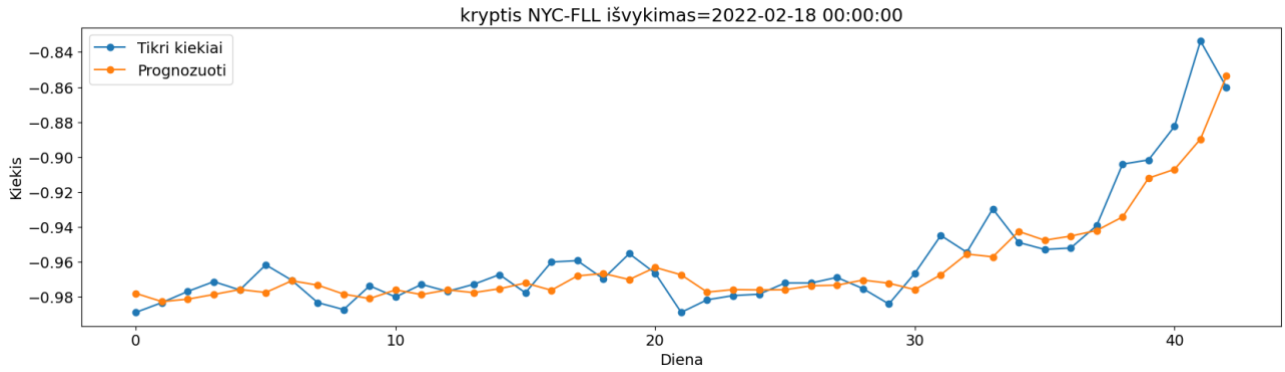


b) Kryptis Niujorkas – Majamis, paieškos artėjant iki išvykimo datos.

35 pav. Seq2Seq kelių oro uostų sekančios dienos prognozės. Treniravimo duomenys. Tikri kiekai (Mėlyna) ir prognozuoti (oranžiniai).



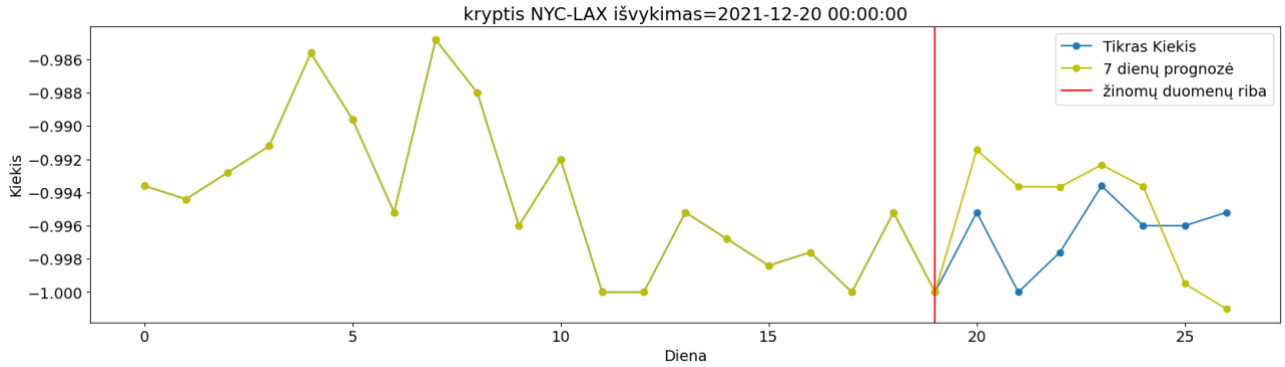
a) Kryptis Majamis – Niujorkas, paieškos artėjant iki išvykimo datos.



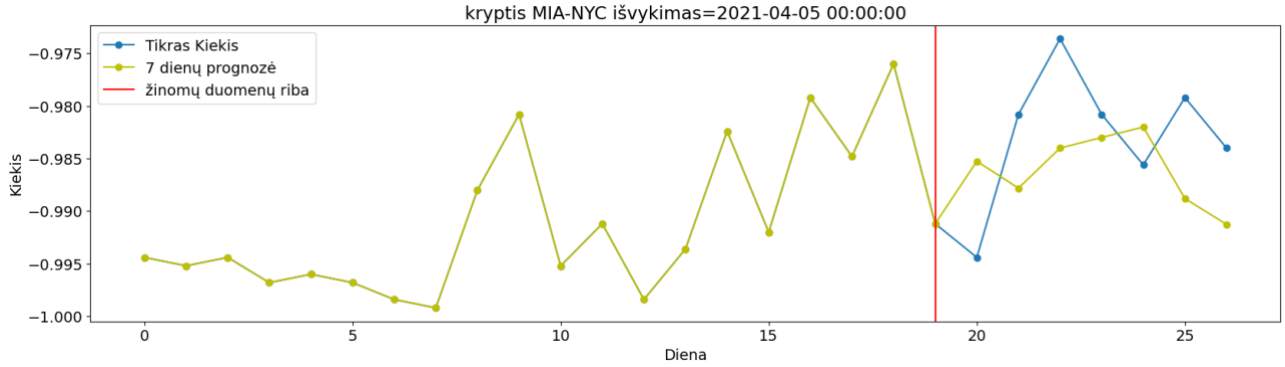
b) Kryptis Niujorkas – Fort Loderdeilas, paieškos artėjant iki išvykimo datos.

36 pav. Seq2Seq kelių oro uostų sekančios dienos prognozės. Testavimo duomenys. Tikri kiekai (Mėlyna) ir prognozuoti (oranžiniai).

Atlikti bandymai su LSTM ir Seq2Seq neuroniniais tinklais rodo labai panašius rezultatus 7 laiko žingsnių uždavinyje. Iš grafikų (žr. 37 pav., 38 pav.) matoma, kad Seq2Seq modelio prognozė ant treniravimo duomenų yra tikslesnė negu LSTM.



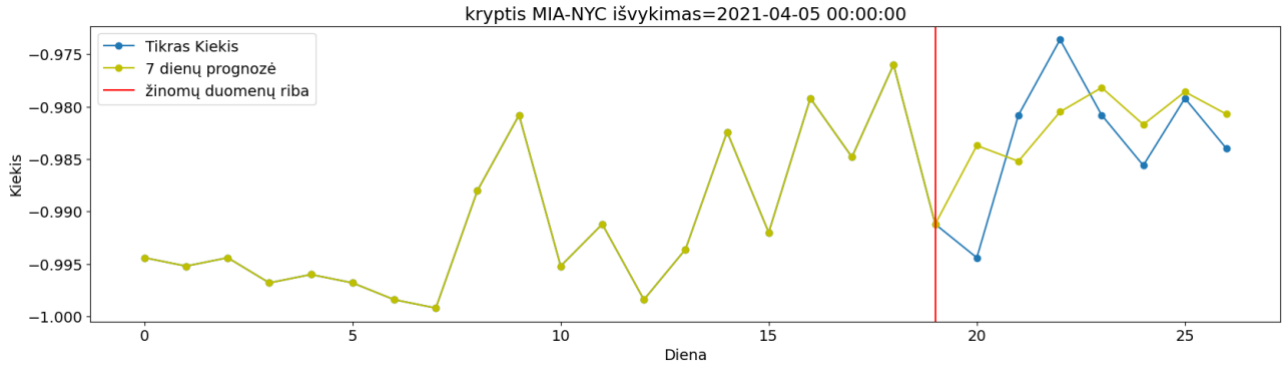
a) Kryptis Niujorkas – Los Andželas.



b) Kryptis Majamis – Niujorkas.

37 pav. LSTM kelių oro uostų sekančių 7 dienų prognozė nuo paskutinių žinomų duomenų taško. Treniravimo duomenys.

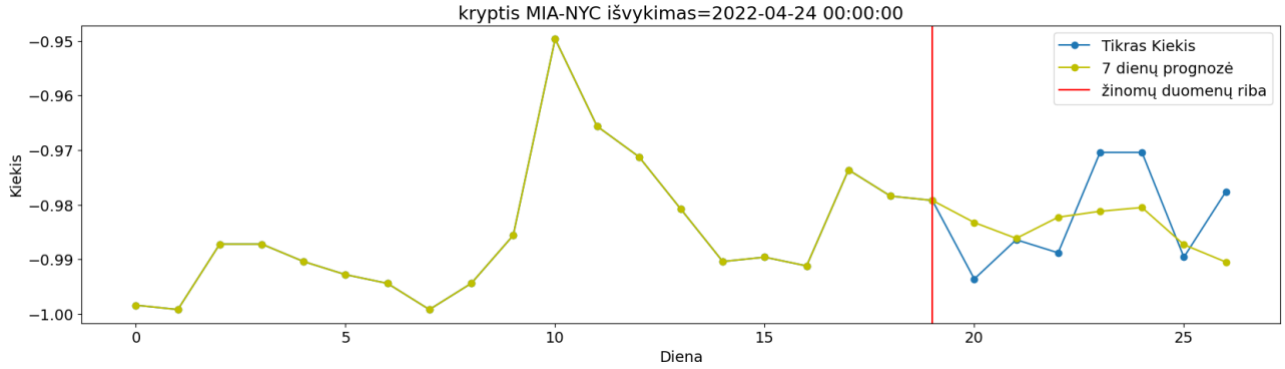
a) Kryptis Niujorkas – Los Andželas.



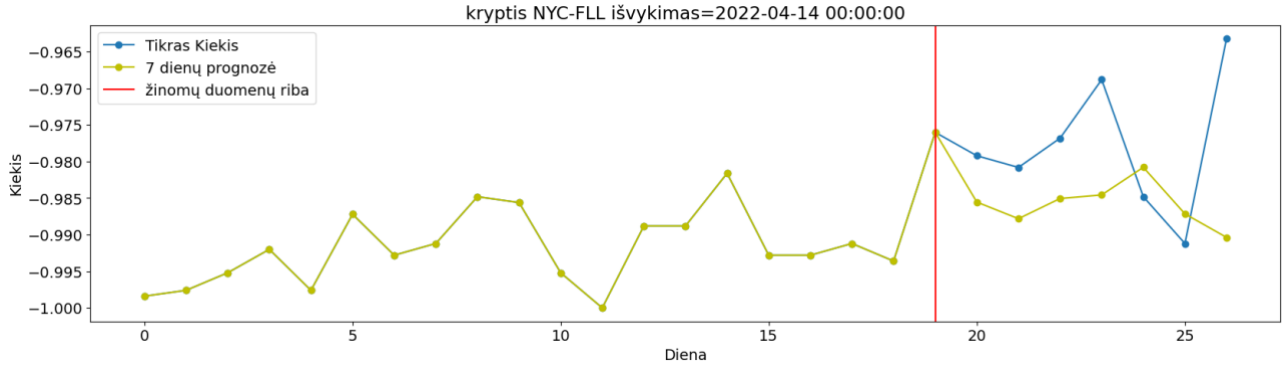
b) Kryptis Majamis – Niujorkas.

38 pav. Seq2Seq kelių oro uostų sekančių 7 dienų prognozė nuo paskutinių žinomų duomenų taško. Treniravimo duomenys.

Kaip ir sekančios dienos prognozės uždavinyje LSTM ir Seq2Seq prognozių rezultatai yra panašūs (žr. 39 pav., 40 pav.). Abiejų modelių prognozės atitinka duomenų tendenciją.

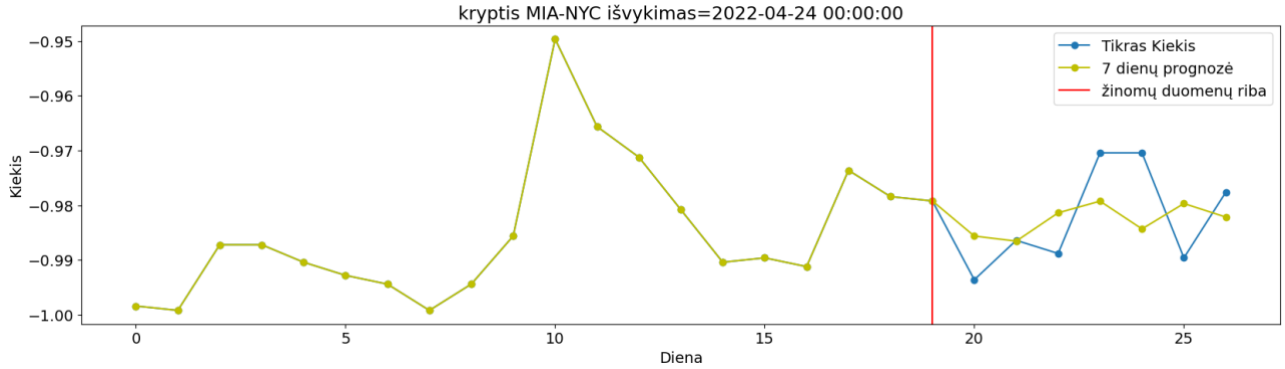


a) Kryptis Majamis – Niujorkas.

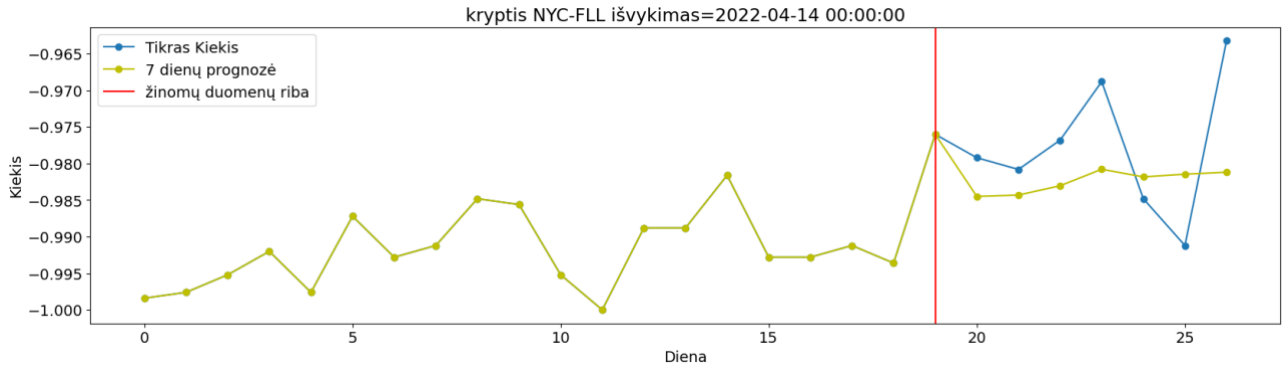


b) Kryptis Niujorkas – Fort Loderdeilas.

39 pav. LSTM kelių oro uostų sekančių 7 dienų prognozė. Testavimo duomenys.



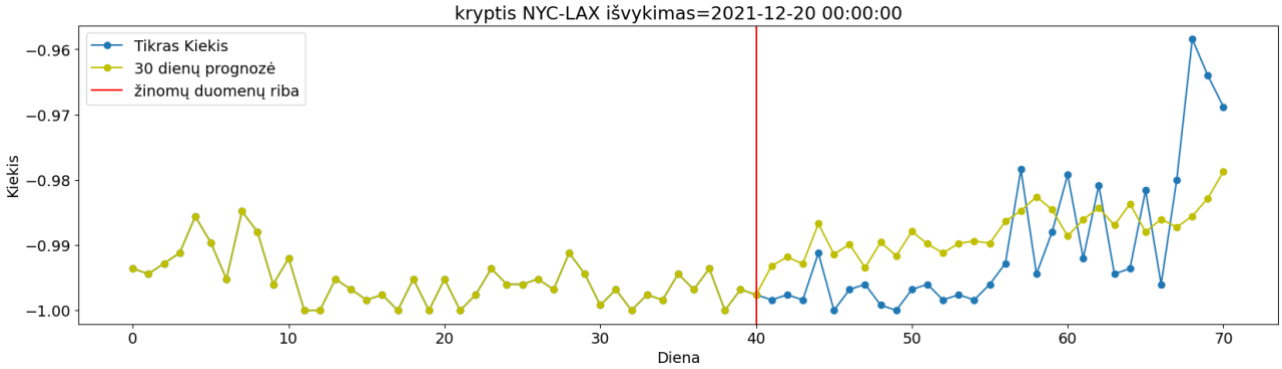
a) Kryptis Niujorkas – Fort Loderdeilas.



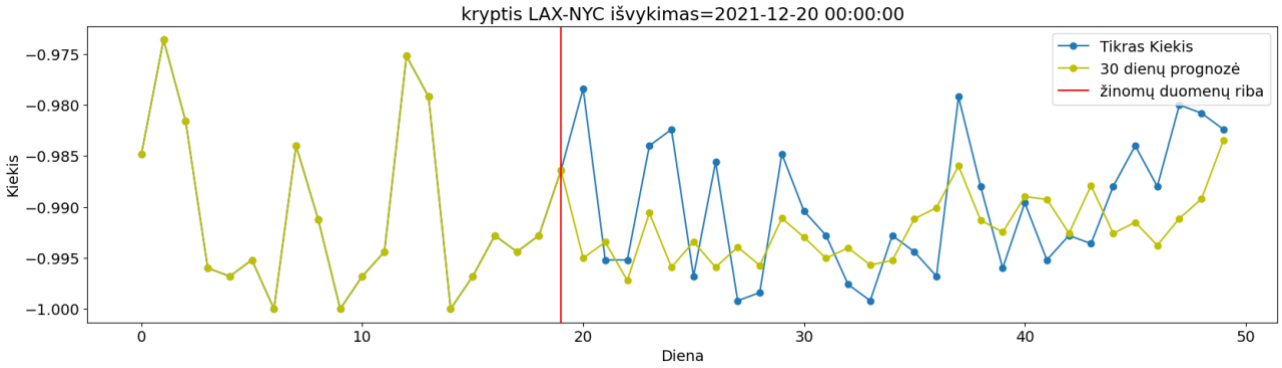
b) Kryptis Niujorkas – Fort Loderdeilas.

40 pav. Seq2Seq kelių oro uostų sekančių 7 dienų prognozė nuo paskutinių žinomų duomenų taško. Testavimo duomenys.

Pastebėjome, kad PDNT neuroninis tinklas netiksliai prognozavo 30 žingsnių į priekį. Atlikti bandymai su sudėtingesniais modeliais: LSTM ir Seq2Seq. Galima matyti treniravimo duomenų prognozių pavyzdžiuose (žr. 41 pav., 42 pav.), kad abiejų modelių prognozavimo tikslumas yra žemas. Tai patvirtina testavimo duomenų prognozės (žr. 43 pav., 44 pav.).

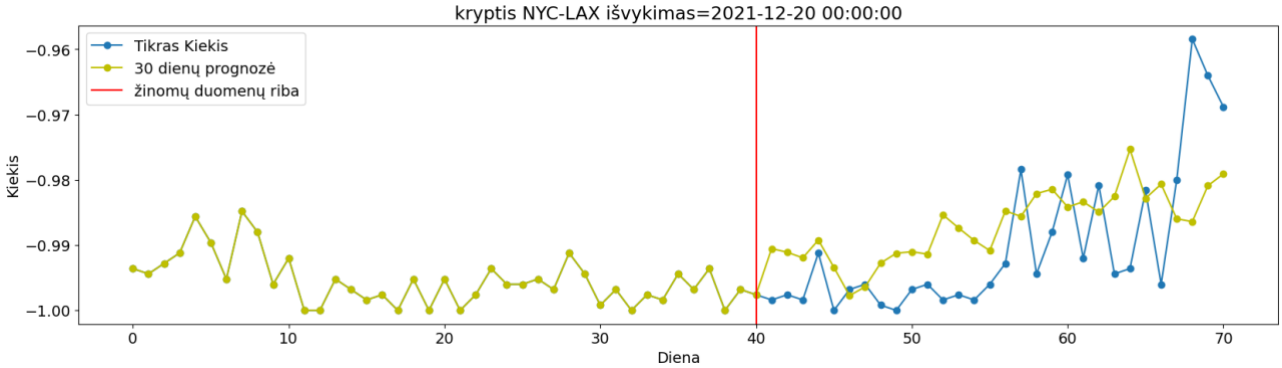


a) Kryptis Niujorkas – Los Andželas.

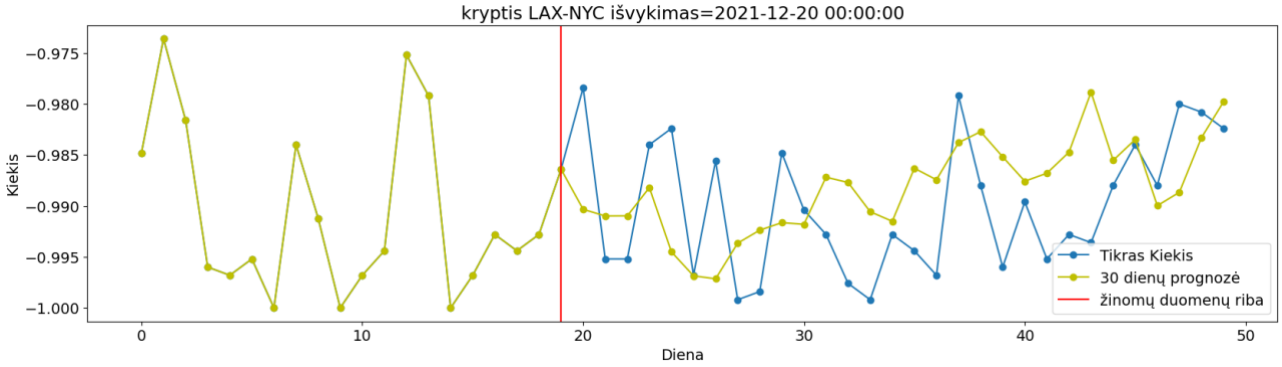


b) Kryptis Los Andželas – Niujorkas.

41 pav. LSTM kelių oro uostų sekančių 30 dienų prognozė nuo paskutinių žinomų duomenų taško. Treniravimo duomenys.

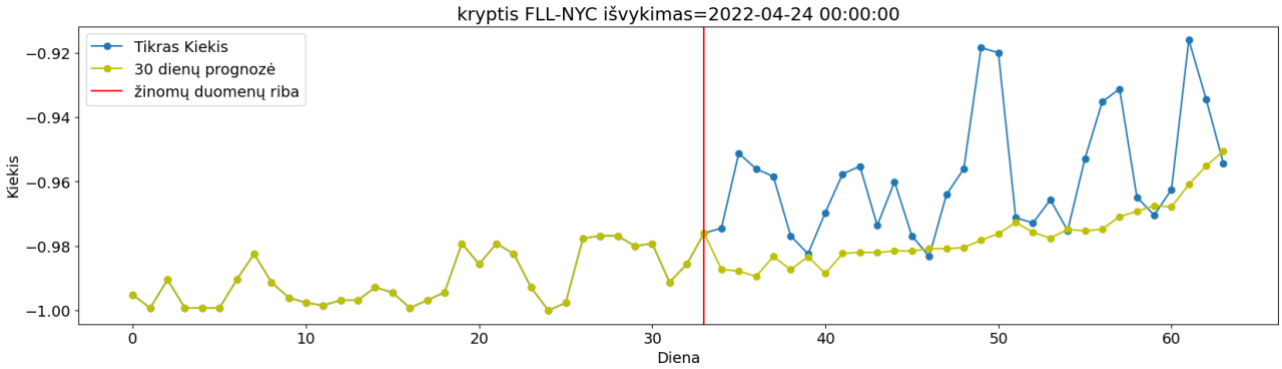


a) Kryptis Niujorkas – Los Andželas.

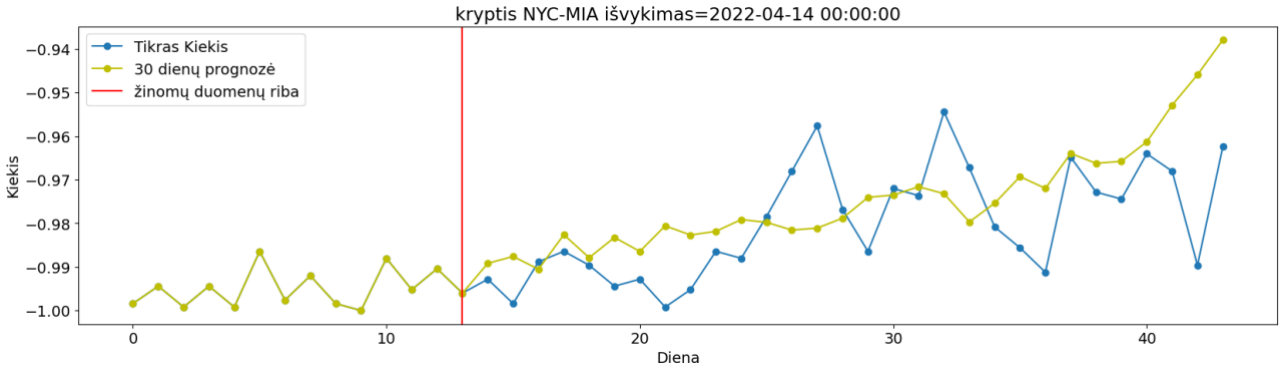


a) Kryptis Los Andželas – Niujorkas.

42 pav. Seq2Seq kelių oro uostų sekančių 30 dienų prognozė nuo paskutinių žinomų duomenų taško. Treniravimo duomenys.

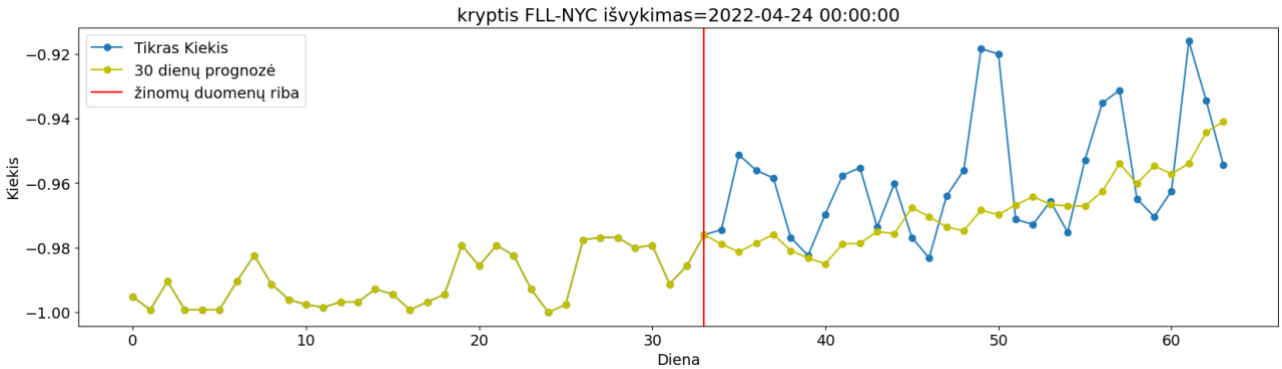


a) Kryptis Fort Loderdeilas – Niujorkas.

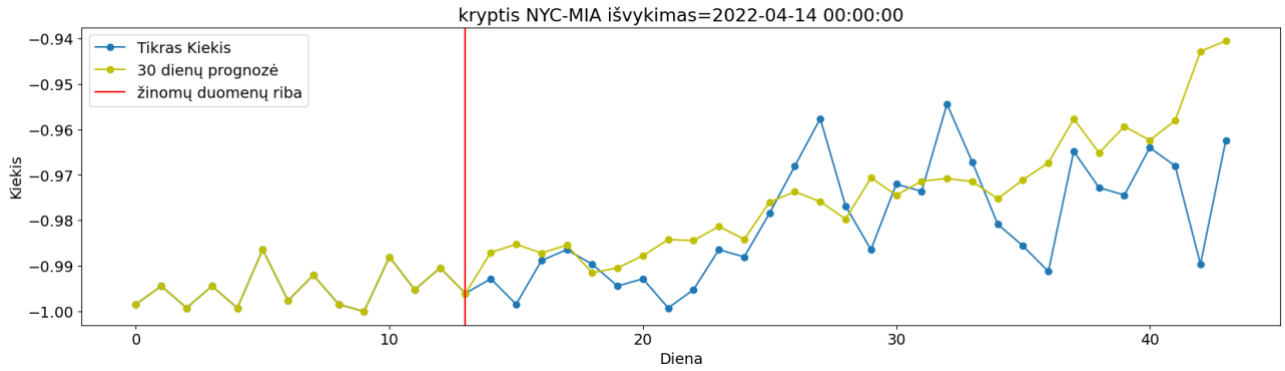


b) Kryptis Niujorkas – Majamis.

43 pav. LSTM kelių oro uostų sekančių 30 dienų prognozė nuo paskutinių žinomų duomenų taško. Testavimo duomenys.



a) Kryptis Fort Loderdeilas – Niujorkas.



b) Kryptis Niujorkas – Majamis.

44 pav. Se2Seq kelių oro uostų sekančių 30 dienų prognozė nuo paskutinių žinomų duomenų taško. Testavimo duomenys.

3.2.4. Skrydžiai į vieną pusę – rezultatai

Atlikti bandymai su trimis neuroninių tinklų modeliais: PDNT, LSTM ir Seq2Seq. Užduotys buvo prognozuoti 1,7 ir 30 laiko žingsnių į priekį. Pateiktose lentelėse matomos oro uostų kryptys, modeliai, vertinimo metrikos ir jų vertės iš tikro duomenų mastelio. Matomos paklaidos kiekvienoje lentelėje apskaičiuotos visai testavimo duomenų imčiai. Mažiausios paklaidos vertės matomos Seq2Seq modelio stulpelyje (žr. **8 lentelė** paryškintos vertės). Svarbu pastebėti, kad modeliai vienu metu buvo mokinami prognozuoti visų pasirinktų oro uostų paieškas, dėl to rezultatai skiriasi kiekvienai krypčiai. Kadangi testavimo duomenyse paieškų kiekių vertės žemos, matomas nedidelis skirtumas tarp visų trijų modelių.

8 lentelė. Skirtingų modelių sekančios dienos prognozės paklaidos iš testavimo duomenų. Vienpusiai skrydžių duomenys.

Sekančios dienos prognozė									
	MAE			**RMSE**			**MAPE**		
Kryptis	**PDNT**	**LSTM**	**Seq2Seq**	**PDNT**	**LSTM**	**Seq2Seq**	**PDNT**	**LSTM**	**Seq2Seq**
MIA-NYC	31.9682	29.3283	**25.1663**	59.7363	54.7517	**46.2167**	25.0070	24.1216	**22.9717**
FLL-NYC	22.8438	21.8282	**19.5436**	37.2316	34.0612	**29.2545**	27.0799	26.5872	**25.4965**
NYC-FLL	14.0609	12.5473	**11.8760**	21.8745	19.9493	**17.4683**	20.6277	**19.3738**	19.9719
NYC-MIA	29.2399	26.1382	**21.2604**	46.3631	42.5060	**35.8502**	26.6696	24.0345	**20.7733**
LAX-NYC	12.0490	10.3313	**9.8668**	17.9313	14.6526	**14.2109**	32.0386	26.8649	**23.7357**
FLL-EWR	7.6199	6.8314	**5.1940**	9.7271	8.9565	**7.8672**	36.1119	31.3119	**19.3202**
LAX-BOS	8.7360	7.3246	**6.9794**	12.5335	**9.7247**	9.7476	42.5797	36.7197	**31.2762**
NYC-LAX	10.8808	10.4112	**10.0588**	16.2683	15.4054	**14.8647**	28.8685	27.9064	**27.6012**
BOS-LAX	6.6063	5.7669	**5.6238**	9.8921	8.9878	8.9282	28.1719	25.7592	**25.6903**
LAX-LAS	7.1415	**6.5730**	7.2489	11.5057	**10.3528**	11.8713	26.1653	**22.2036**	25.1146
LAS-LAX	10.1734	9.3192	**9.1605**	13.4765	12.1659	**12.1642**	34.0444	**27.2591**	28.7174
LAX-SFO	6.2126	5.4640	**5.0096**	7.8086	8.1497	**7.5638**	27.6670	**20.1688**	20.3691

SFO-NYC	11.4910	**6.7502**	7.4275	15.7833	**9.8388**	10.5806	40.2208	**23.1246**	23.2973
LAX-MIA	11.1238	16.2032	**11.2908**	14.4723	19.7188	**14.7531**	33.2251	55.8889	**38.3513**
SFO-LAX	5.2489	6.0687	**4.9783**	7.6870	8.4071	**7.1647**	27.1944	26.4183	**20.4833**
ATL-NYC	13.9593	9.9261	**8.8095**	15.9718	**15.4546**	15.6891	65.0265	36.6852	**26.0247**
NYC-ATL	9.6309	9.1048	**9.0436**	12.6480	12.1735	**11.4346**	32.7420	30.1871	**29.9961**

Lentelėse (žr.

9 lentelė, **10 lentelė**) matomos paklaidos prognozuojant 7 ir 30 laiko žingsnių į priekį. Vertinimo metrikos rodo, kad LSTM ir Seq2Seq modeliai daro tiksliausias 7 laiko žingsnių prognozes (žr.

9 lentelė), tačiau pažvelgus į 30 laiko žingsnių prognozių paklaidas (žr. **10 lentelė**), matoma, kad visų modelių paklaidos yra beveik vienodos. Aukštos paklaidos **10 lentelė** patvirtina, kad modeliai nėra pilnai išmokę ilgalaikių ryšių reikalingų 30 dienų prognozėms.

9 lentelė. Skirtingų modelių sekančių 7 dienų prognozių paklaidos iš testavimo duomenų. Vienpusiai skrydžių duomenys.

Sekančių 7 dienų prognozė									
	MAE			**RMSE**			**MAPE**		
Kryptis	**PDNT**	**LSTM**	**Seq2Seq**	**PDNT**	**LSTM**	**Seq2Seq**	**PDNT**	**LSTM**	**Seq2Seq**
MIA-NYC	33.8919	27.0321	21.9898	46.3404	36.2487	30.4291	43.8806	40.4082	32.9894
FLL-NYC	23.7378	21.5779	16.3823	32.8435	29.6391	23.6374	40.9023	43.7036	34.8441
NYC-MIA	18.4972	16.8662	14.9454	27.9986	25.7544	23.6598	37.5312	37.9371	39.2997
NYC-FLL	15.5516	12.8435	11.2678	20.1209	17.4289	15.4217	38.3920	32.0822	32.7470
NYC-LAX	15.1273	15.2611	11.3920	20.3674	20.0457	15.9401	37.0402	37.0492	29.6577
LAX-NYC	13.8883	12.5192	9.7163	18.8871	16.5468	13.4596	37.4392	34.0766	29.0864
FLL-EWR	9.4469	10.4581	11.4924	13.0352	16.5291	16.8651	38.4425	43.8798	54.3130
LAX-LAS	9.0234	7.4995	9.3933	17.3032	15.4716	16.2951	40.8714	36.5497	51.4374
LAX-BOS	5.4139	5.3394	6.5184	6.8067	6.3627	7.7530	31.8041	33.8773	43.8016
BOS-LAX	6.8777	6.6060	6.6662	8.4262	8.1721	8.1445	36.1584	35.8138	42.8516
NYC-SFO	6.3003	6.7566	4.5520	7.7693	8.6407	5.6122	32.5668	33.8644	24.3020
LAS-LAX	8.3475	6.2497	7.1002	10.8505	8.1862	8.5730	38.3833	28.9764	39.7340
ATL-NYC	9.1093	8.4848	7.6585	12.3690	10.5577	9.5270	31.2108	31.2416	28.6417
SFO-NYC	6.8625	6.9288	5.6727	8.5816	8.9590	7.0958	35.8606	34.9754	32.2700

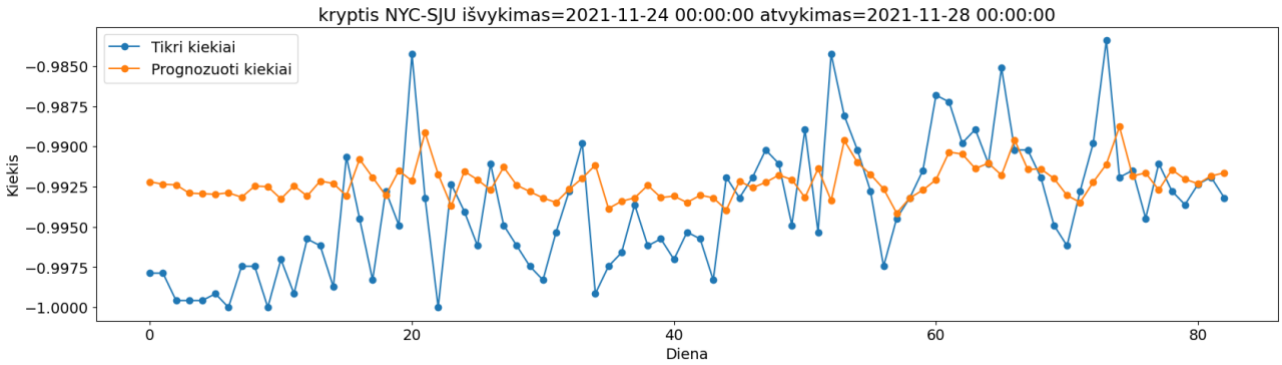
10 lentelė. Skirtingų modelių sekančių 30 dienų prognozių paklaidos iš validacijos duomenų. Vienpusiai skrydžių duomenys.

Sekančių 30 dienų prognozė									
	MAE			**RMSE**			**MAPE**		
Kryptis	**PDNT**	**LSTM**	**Seq2Seq**	**PDNT**	**LSTM**	**Seq2Seq**	**PDNT**	**LSTM**	**Seq2Seq**
MIA-NYC	27.1428	27.8290	26.4959	38.9756	38.8541	36.6940	42.5717	39.9503	38.4594

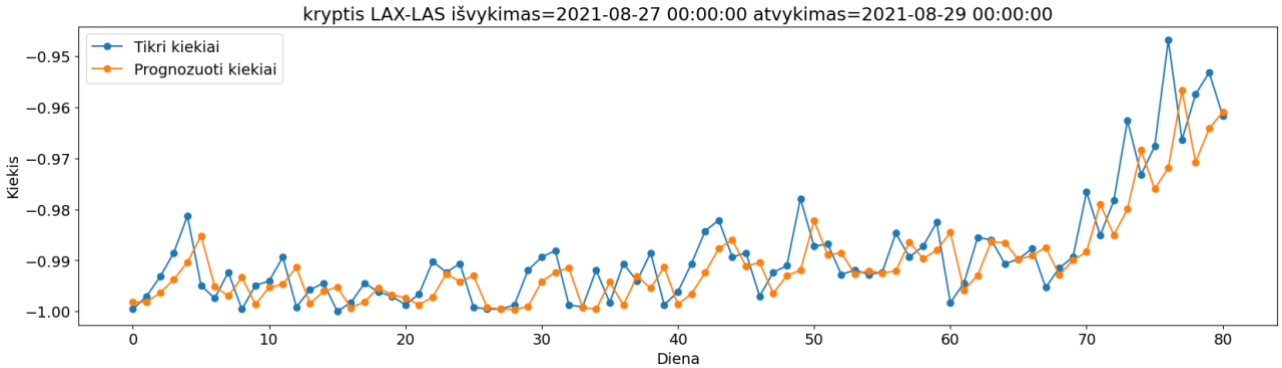
FLL-NYC	19.1660	19.4953	18.1920	26.5028	27.2960	25.8528	45.0220	36.8141	34.8480
NYC-MIA	20.7126	13.9888	12.8526	27.5380	21.1651	18.6467	73.4743	40.8567	39.9315
NYC-FLL	19.0785	14.1534	13.0897	24.1458	18.9831	16.6192	68.9217	47.1745	44.0299
NYC-LAX	22.1393	10.5310	9.8475	26.1777	13.9664	12.7423	58.3572	29.6731	31.6776
LAX-NYC	11.2968	13.1312	13.8461	15.0715	16.6216	17.3902	33.7120	50.0217	53.3003
FLL-EWR	20.1990	15.5351	15.9472	22.2949	18.0876	18.1419	110.8532	83.2021	85.2647
LAX-LAS	18.2009	16.2680	18.2239	26.0961	22.6297	25.4769	86.9582	91.7875	96.2569
LAX-BOS	7.9480	23.2561	23.0762	10.0244	24.4939	23.8013	49.4359	145.1357	144.8478
BOS-LAX	14.3063	13.4105	16.1162	16.7715	15.9357	17.9112	78.8383	84.6798	100.0349
NYC-SFO	10.8861	5.8761	5.3930	12.2877	7.6298	6.7157	55.6441	33.8540	31.1145
LAS-LAX	27.1589	15.0825	10.5380	35.4061	19.1035	14.0024	97.1566	67.4223	49.7171

3.2.5. Skrydžiai į abi puses

Dvipusiai skrydžiai sudaro didelę dalį kelionės paieškų. Šis skrydžių tipas turi dar vieną parametrą – atvykimo datą. Dėl šio papildomo parametro skrydžių variacijos kiekis stipriai padidėja. Kiekviena dvipusių skrydžių kryptis turi 3 datas: paieškos, išvykimo ir atvykimo. Šie duomenys turi skirtingus ryšius ir didesnės imties sekas (žr. 45 pav.), negu vienpusiai skrydžiai. Atliekami identiški tyrimai, daryti su vienpusių skrydžių duomenimis, atsižvelgiant į dvipusių skrydžių duomenų ryšius.



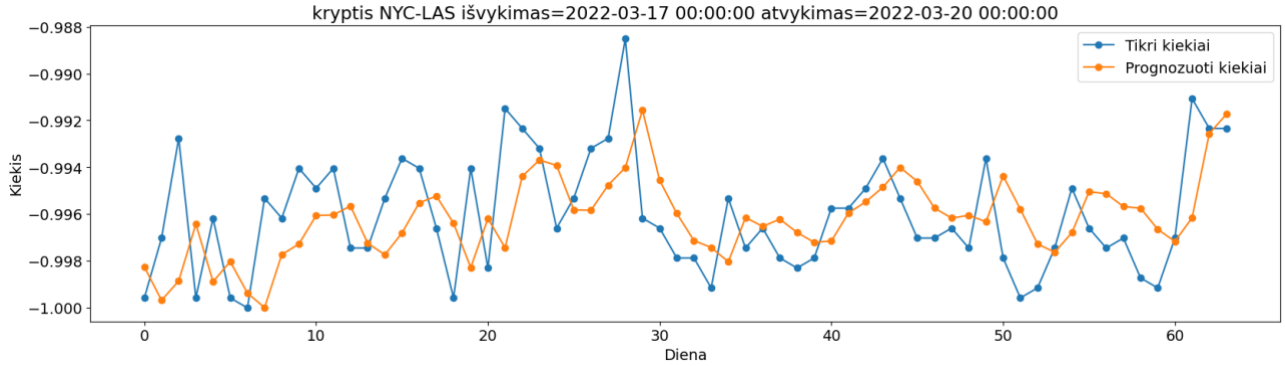
a) Kryptis Niujorkas – Luis Munoz Marin, paieškos artėjant iki išvykimo datos.



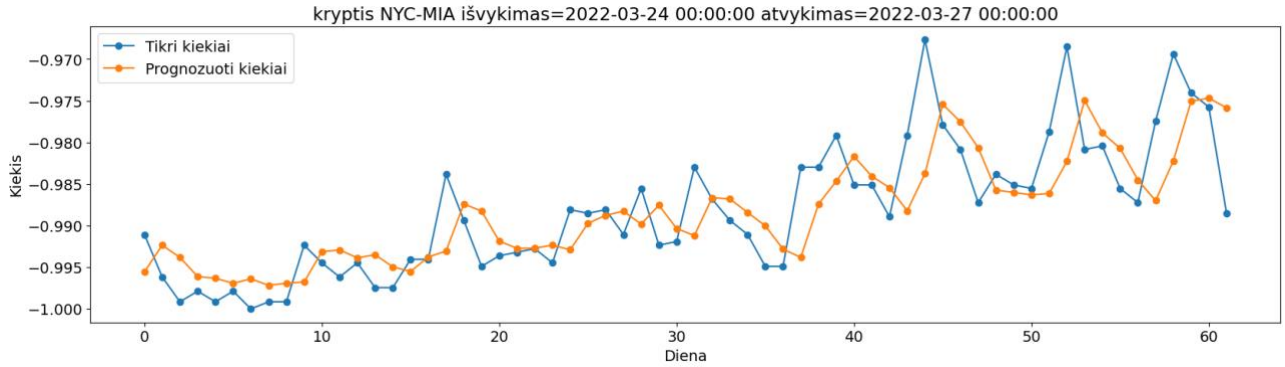
b) Kryptis Los Andželas – Las Vegasas, paieškos artėjant iki išvykimo datos.

45 pav. PDNT kelių oro uostų sekančios dienos prognozės. Treniravimo duomenys. Tikri duomenys (mėlyna) ir prognozuoti (oranžinė).

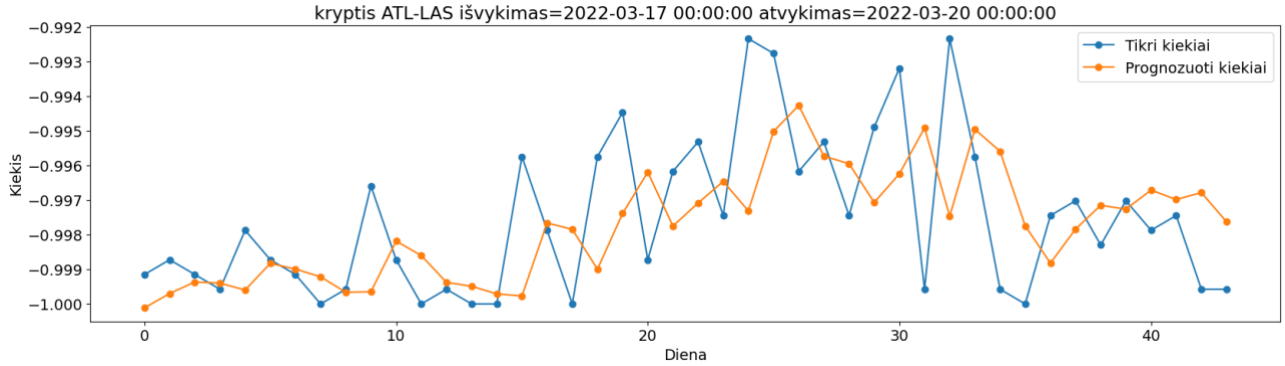
Kaip ir su vienpusiais skrydžių duomenimis, LSTM (žr. 46 pav.) ir Seq2Seq (žr. 47 pav.) modelių prognozės rezultatai yra netoli tikrų paieškų verčių.



a) Kryptis Niujorkas – Las Vegasas, paieškos artėjant iki išvykimo datos.

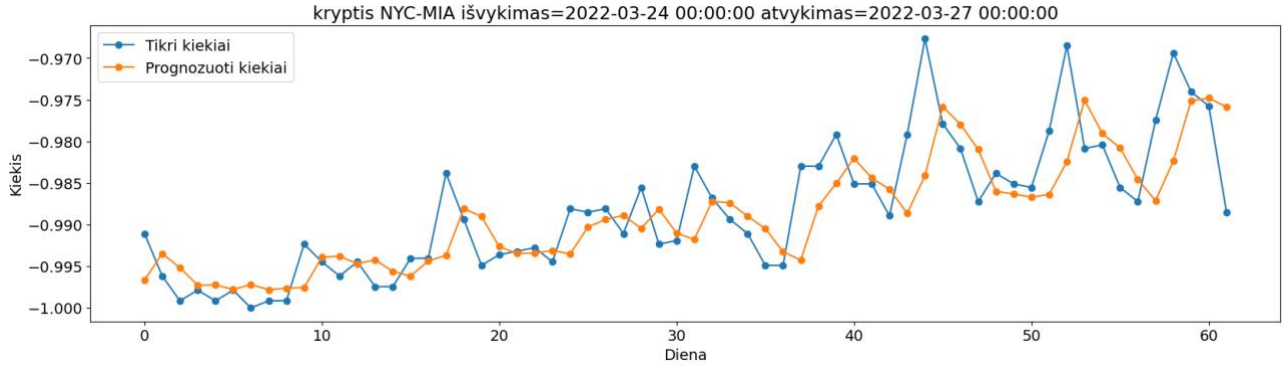


b) Kryptis Niujorkas – Majamis, paieškos artėjant iki išvykimo datos.

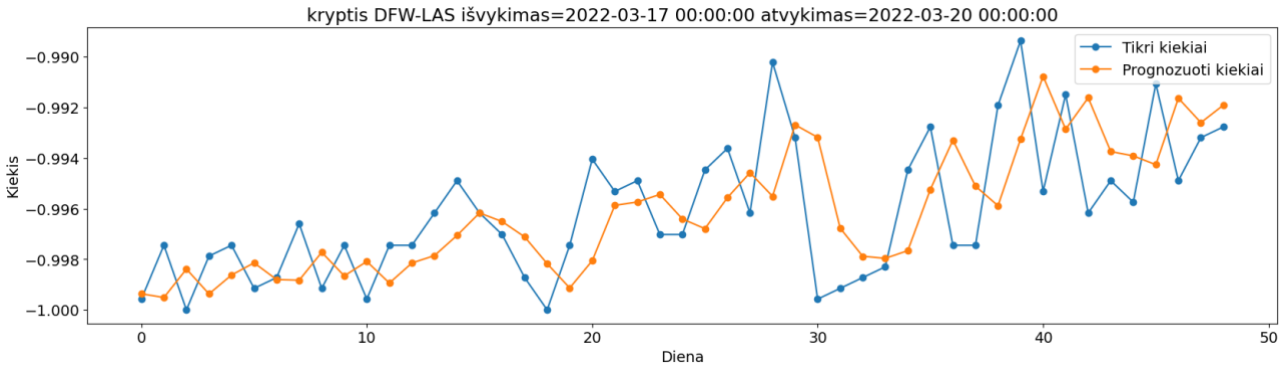


c) Kryptis Atlanta – Las Vegasas, paieškos artėjant iki išvykimo datos.

46 pav. LSTM kelių oro uostų sekančios dienos prognozės. Testavimo duomenys. Tikri duomenys (mėlyna) ir prognozuoti (oranžinė).



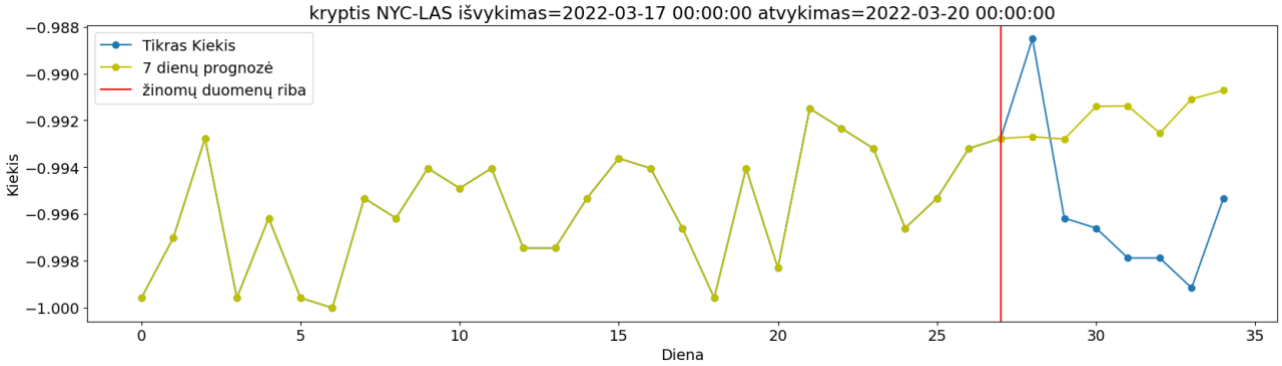
a) Kryptis Niujorkas – Majamis, paieškos artėjant iki išvykimo datos.



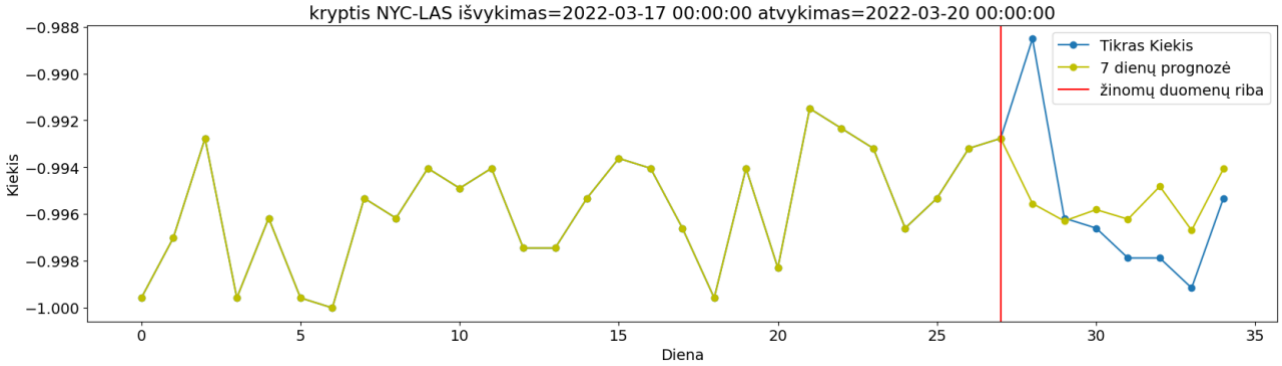
b) Kryptis Dalasas – Las Vegasas, paieškos artėjant iki išvykimo datos.

47 pav. Seq2Seq modelio sekančios dienos prognozės iš testavimo duomenų. Tikri duomenys (mėlyna) ir prognozuoti (oranžinė).

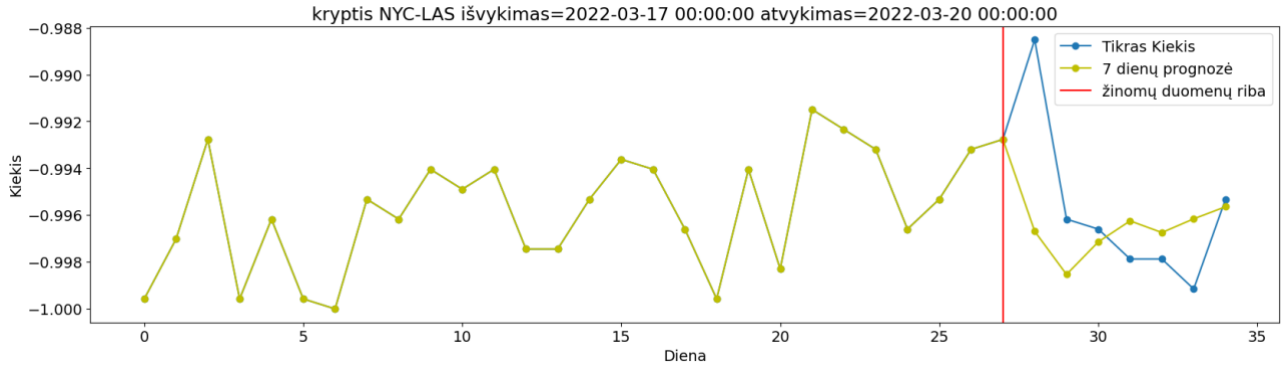
Skrydžiai į abi puses yra aktuali skrydžių paieškų dalis. Vartotojai atidžiau, bet dažniau renkasi šio tipo paieškas siekdami sutaupyti laiką, gauti geresnių pasiūlymų ar lengviau suderinti apgyvendinimo klausimus. Dėl šių priežasčių, vartotojai paieškas daro iš anksčiau, palengva planuodami savo kelionę. Atliekami tyrimai kaip šiuos vartotojų elgesio bruožus pastebės treniruojami modeliai: PDNT, LSTM ir Seq2Seq. Pateiktuose modelių prognozių rezultatuose (žr. 48 pav., 49 pav., 50 pav.), matoma, kad LSTM modelio prognozė yra artimiausia tikrųjų verčių.



48 pav. PDNT septynių dienų prognozė iš testavimo duomenų nuo paskutinių žinomų duomenų taško.

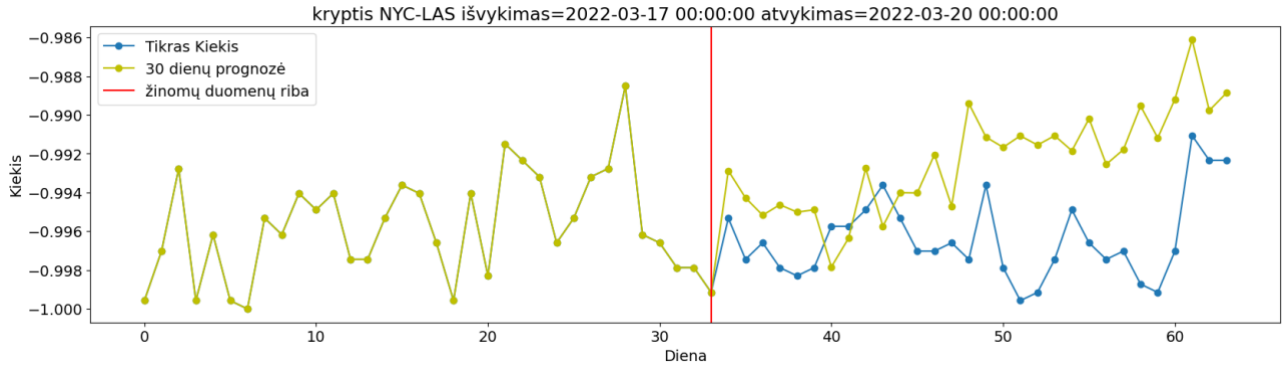


49 pav. LSTM septynių dienų prognozė iš testavimo duomenų nuo paskutinių žinomų duomenų taško.

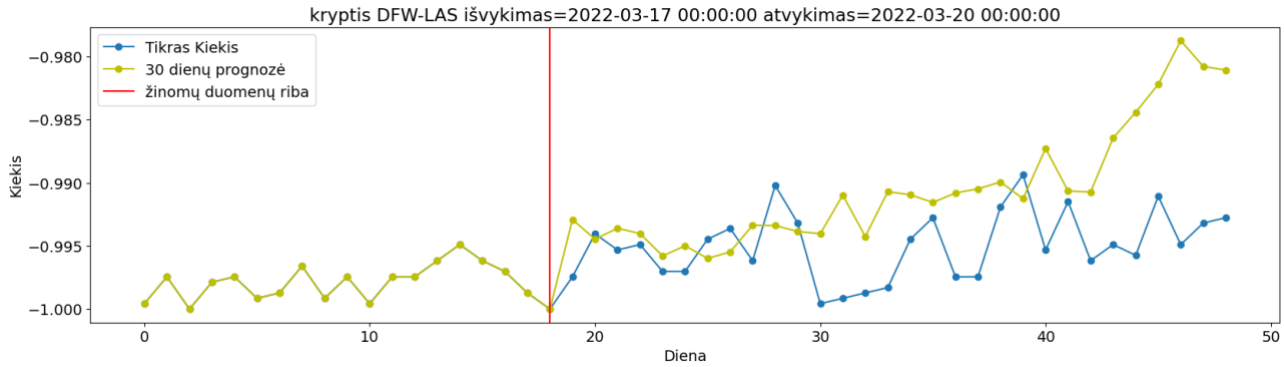


50 pav. Seq2Seq septynių dienų prognozė iš testavimo duomenų nuo paskutinių žinomų duomenų taško.

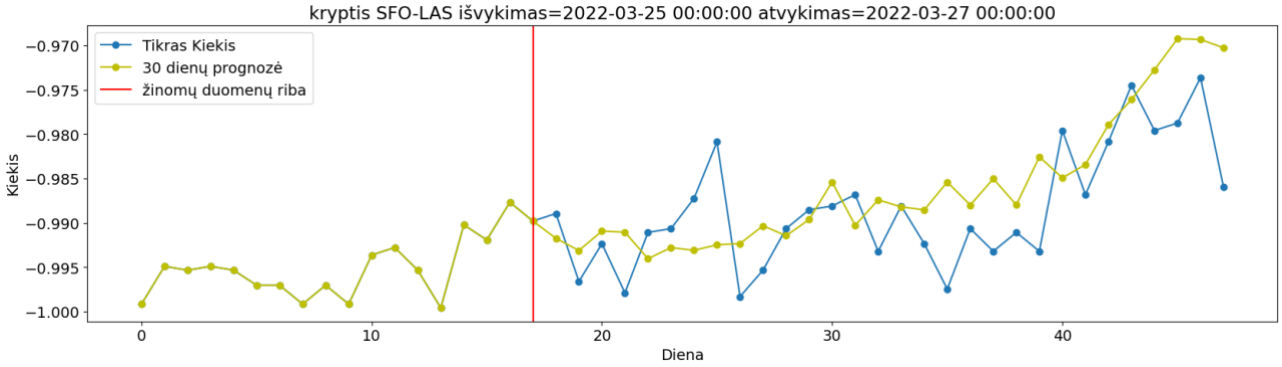
Prognozių pavyzdžiuose pastebima, PDNT modelio prognozuotos vertės (žr. 51 pav.) nukrypsta nuo tikrų duomenų, tuo tarpu LSTM (žr. 52 pav.) ir ypatingai Seq2Seq (žr. 53 pav.) prognozės yra arčiau tikrųjų paieškos verčių. Tai patvirtina, kad ilgos dvipusių skrydžių duomenų sekos geriau įsisavinamos modelių su rekurentiniais sluoksniais.



a) Kryptis Niujorkas– Las Vegasas.

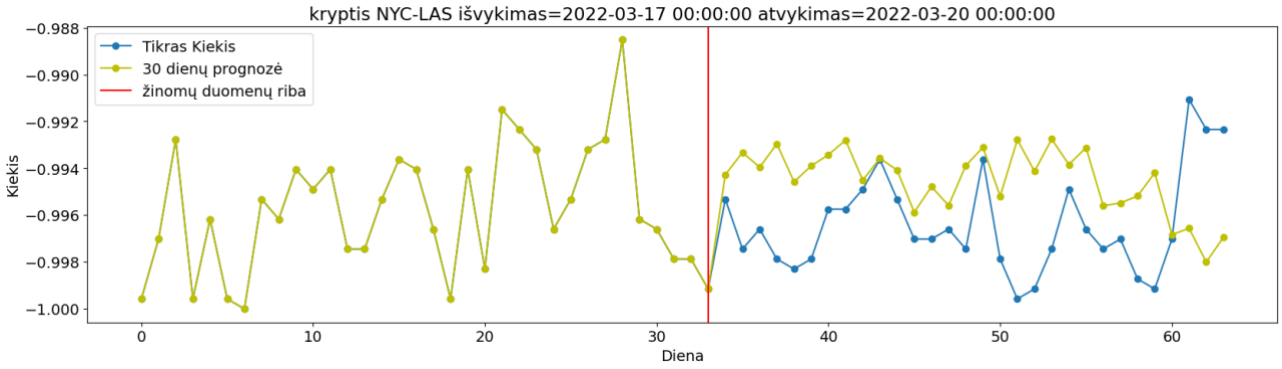


b) Kryptis Dalasas – Las Vegasas.

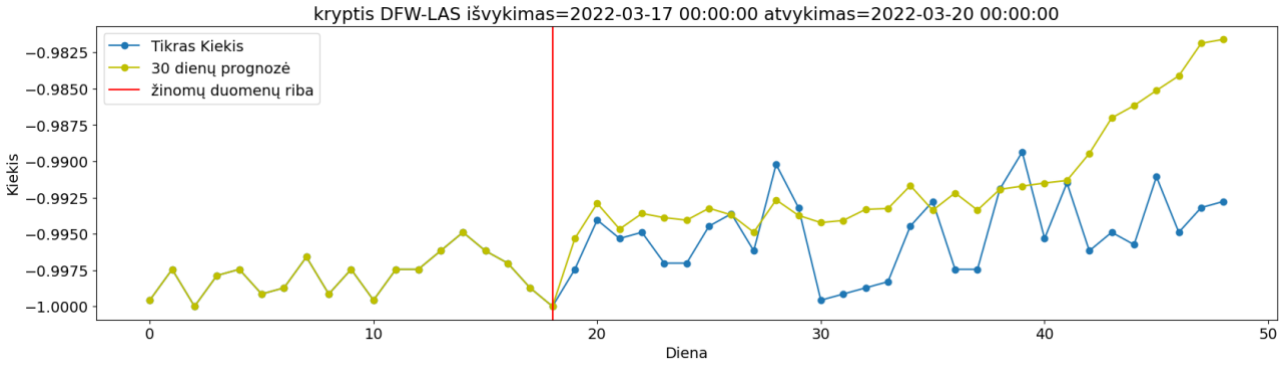


c) Kryptis San Franciskas – Las Vegasas.

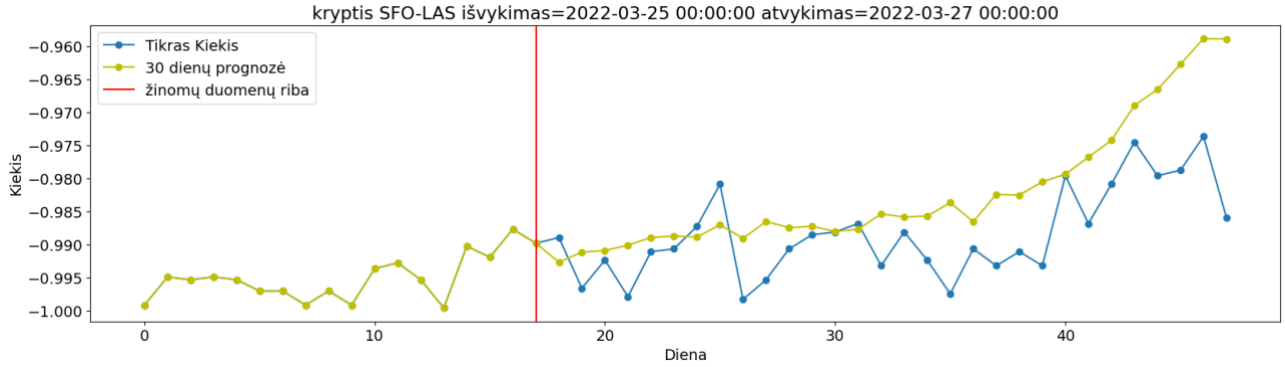
51 pav. PDNT kelių oro uostų 30 dienų prognozės nuo paskutinių žinomų duomenų taško. Testavimo duomenys.



a) Kryptis Niujorkas – Las Vegasas.

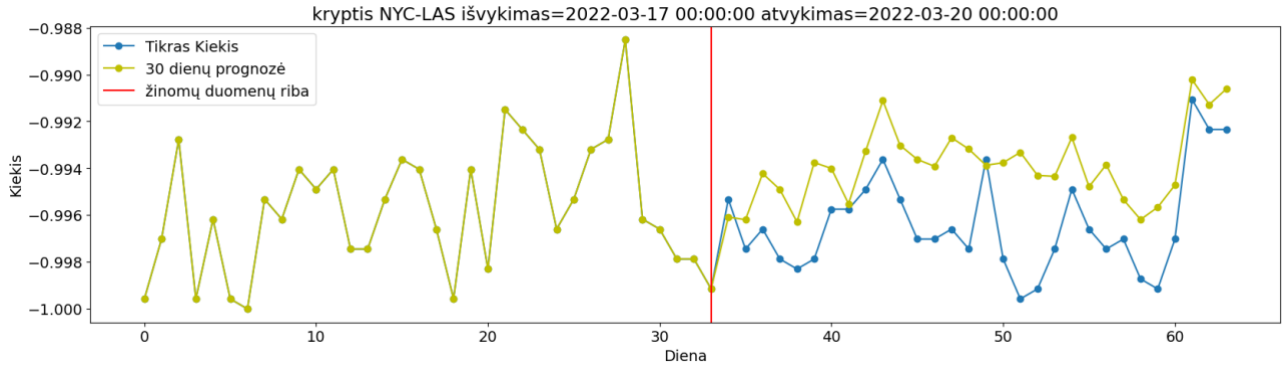


b) Kryptis Dalasas– Las Vegasas.

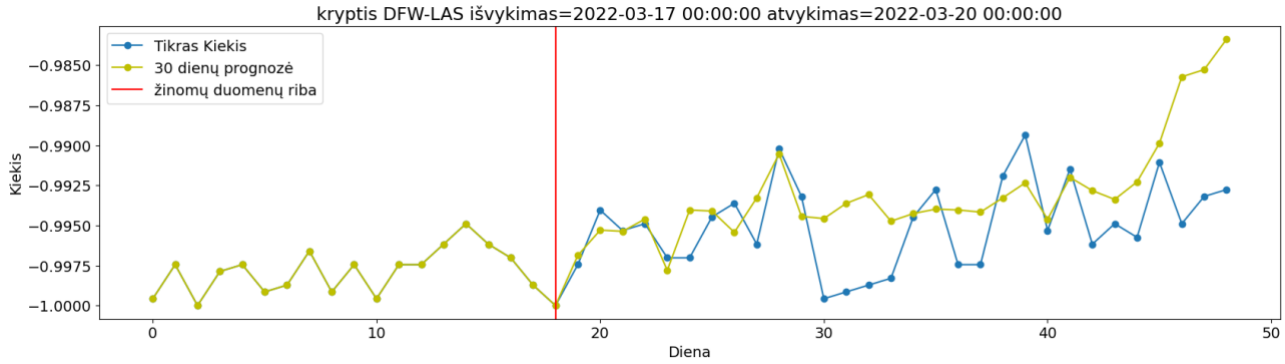


c) Kryptis San Franciskas – Las Vegasas.

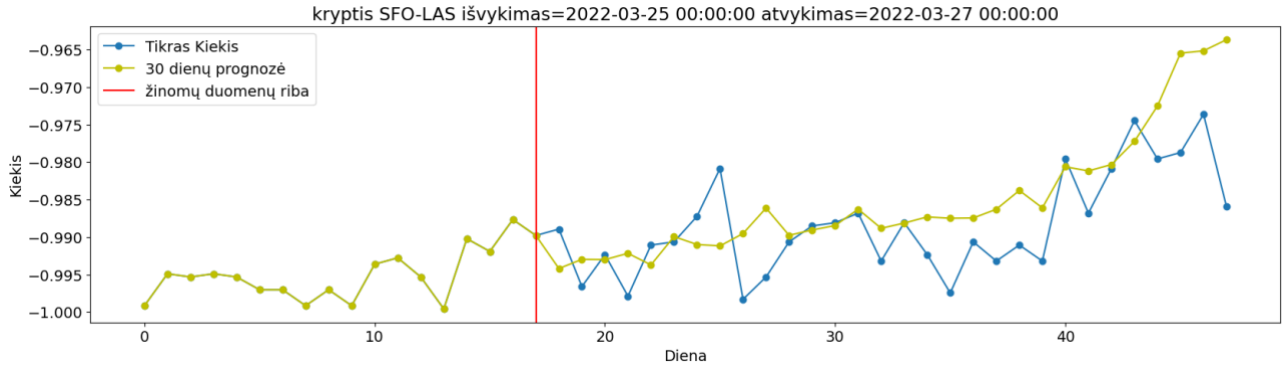
52 pav. LSTM kelių oro uostų 30 dienų prognozės nuo paskutinių žinomų duomenų taško. Testavimo duomenys.



a) Kryptis Niujorkas – Las Vegasas.



b) Kryptis Dalasas – Las Vegasas.



c) Kryptis San Franciskas – Las Vegasas.

53 pav. Seq2Seq kelių oro uostų 30 dienų prognozės nuo paskutinių žinomų duomenų taško. Testavimo duomenys.

3.2.6. Skrydžiai į abi puses - rezultatai

Atlikti bandymai su dvipusių skrydžių duomenimis, apmokinant tris neuroninio tinklo architektūras: PDNT, LSTM ir Seq2Seq. Užduotys buvo prognozuoti 1,7 ir 30 laiko žingsnių į priekį. Žemiau esančiose lentelėse yra pateiktos prognozavimo paklaidos, kurios apskaičiuotos visai dvipusių skrydžių treniravimo duomenų imčiai. Paklaidoms skaičiuoti naudojami duomenys yra tikro mastelio (t. y. neminimizuoti ir nenormalizuoti). Sekančios dienos prognozės paklaidos tarp visų modelių išlieka panašios per visus oro uostų derinius (žr. **11 lentelė**).

11 lentelė. Skirtingų modelių sekančios dienos prognozės paklaidos iš testavimo duomenų. Dvipusiai skrydžių duomenys.

Sekančios dienos prognozė									
	MAE			**RMSE**			**MAPE**		
Kryptis	**PDNT**	**LSTM**	**Seq2Seq**	**PDNT**	**LSTM**	**Seq2Seq**	**PDNT**	**LSTM**	**Seq2Seq**
NYC-LAS	5.1297	4.8627	5.0879	6.2106	5.9095	6.2720	30.7582	24.9697	25.1520
NYC-MIA	9.1307	9.5106	9.1270	11.9158	12.7023	12.7156	28.0842	26.6683	23.8081
DFW-LAS	4.9581	4.5228	4.6179	6.2546	5.7288	5.7052	23.6527	22.7391	23.9433
SFO-LAS	9.2634	8.8374	9.0505	11.6749	11.4201	11.3731	34.8124	34.8022	33.8492
LAX-LAS	11.5514	11.4888	11.6118	18.5073	17.3226	17.4064	24.3955	25.1530	29.8846
ATL-LAS	4.7377	3.9314	4.1597	6.3595	5.0972	5.2859	26.3067	23.4081	25.7545

Kitokia situacija matoma kelių laiko žingsnių prognozės rezultatuose. Matome daugelyje oro uostų derinių PDNT modelio paklaidos ženkliai didesnės negu LSTM ir Seq2Seq (žr. **12 lentelė**, **13 lentelė**).

12 lentelė. Skirtingų modelių sekančių 7 dienų prognozių paklaidos iš testavimo duomenų. Dvipusiai skrydžių duomenys.

7 dienų prognozė									
	MAE			**RMSE**			**MAPE**		
Kryptis	**PDNT**	**LSTM**	**Seq2Seq**	**PDNT**	**LSTM**	**Seq2Seq**	**PDNT**	**LSTM**	**Seq2Seq**
NYC-LAS	13.4445	8.1790	7.5766	15.6467	10.9765	10.1704	79.3299	44.0046	38.9249
NYC-MIA	16.0621	11.4926	11.1730	19.5667	15.1708	15.3768	56.2056	35.6288	28.5481
DFW-LAS	8.7663	6.1429	5.7435	10.1535	8.2173	7.4762	52.1451	30.4527	28.7486
SFO-LAS	9.7814	8.8679	9.6725	11.9082	10.8796	11.9857	43.8633	30.7812	31.1449

13 lentelė. Skirtingų modelių sekančių 30 dienų prognozių paklaidos iš testavimo duomenų. Dvipusiai skrydžių duomenys.

30 dienų prognozė			
	MAE	**RMSE**	**MAPE**

Kryptis	**PDNT**	**LSTM**	**Seq2Seq**	**PDNT**	**LSTM**	**Seq2Seq**	**PDNT**	**LSTM**	**Seq2Seq**
NYC-LAS	8.4878	6.0447	5.9598	10.1607	7.3926	7.3458	48.2828	33.1852	32.1345
NYC-MIA	10.0047	9.2506	9.6690	12.8613	11.9938	12.8011	30.6390	26.6680	25.7976
DFW-LAS	8.7313	7.5142	6.0422	11.1691	9.3015	7.7734	51.5405	44.6451	32.8525
SFO-LAS	12.6090	12.8291	11.2449	15.7811	16.9796	15.1074	53.5890	52.4251	44.2154

Išvados

1. Darbe ištirti skirtingi statistiniai ir dirbtinio intelekto modeliai kelionių paieškų kiekio prognozavimui ir išskirti potencialūs modelio patobulinimai atsižvelgiant į uždavinio specifiką ir duomenų struktūrą.
2. Geografinių koordinačių panaudojimas oro uostų užkodavimo tikslais leido apmokyti dirbtinio neuroninio tinklo modelius atskirti skirtumus ir panašumus tarp oro uostų. Toks sprendimas leido modeliams efektyviau prognozuoti artimų objektų vertes, net ir turint ribotą duomenų kiekį.
3. Ekspertinės analizės būdu nustatyti vartotojų paieškos elgsenos šablonai, siekiant įsisavinti vartotojų elgsenos tendencijas, kurios kartojasi kiekvieną mėnesį, leido atrinkti tinkamiausias modelių charakteristikas.
4. Atlikti eksperimentų rezultatai, parodė, kad tiesinis neuroninis tinklas apmokytas prognozuoti pagal išvykimo datą, pateikia žemo tikslumo skrydžių paieškų kiekio prognozavimo rezultatus su paklaidomis 462.66 RMSE ir 74.01% MAPE. Tuo tarpu ARIMA modelis parodė ženkliai aukštesnius prognozavimo rezultatus su paklaida 362.16 RMSE ir MAPE 56.72%. Aukščiausias tikslumas gautas su LSTM ir Seq2Seq modeliais, iš kurių Seq2Seq autoenkoderis parodė pranašumą su mažiausia paklaida siekiančią 267.85 RMSE ir 52.38% MAPE.
5. Sukurtas sprendimas, kuriame įtraukti du paieškų kiekių prognozavimo kriterijai: (1) pagal skrydžio išvykimo datą, ir (2) pagal skrydžio paieškos datą. Prognozavimo pagal paieškos datą eksperimentuose, PDNT, LSTM ir Seq2Seq autoenkoderio sekančios dienos prognozavimo rezultatai ženkliai nesiskiria, tačiau daugeliu atvejų Seq2Seq modelio prognozės skirtingiems oro uostų deriniams buvo tiksliausios, lenkiant PDNT modelį vidutiniškai 3-15% ir LSTM 1-10% vertinant MAPE metriką. Kelių laiko žingsnių prognozavimo uždaviniuose taikomi LSTM ir Seq2Seq modeliai parodė pranašumą, 7 ir 30 laiko žingsnių (dienų) prognozėse pagerindami PDNT modelio tikslumą vidutiniškai 10-30% vertinant MAPE metriką.

Literatūros sąrašas

1. BANDARA, K. ir kt. Sales Demand Forecast in E-commerce using a Long Short-Term Memory Neural Network Methodology. [interaktyvus]. .arXiv, 2019arXiv:1901.04028 [cs, stat]. . Prieiga per: http://arxiv.org/abs/1901.04028

2. WANG, L. ir kt. Deep Learning for Flight Demand Forecasting. . . Prieiga per: doi: https://arxiv.org/pdf/2011.04476.

3. SCHAER, O. ir kt. Demand forecasting with user-generated online information. *International Journal of Forecasting*. 2019, 35(1), 197–212. Prieiga per: doi: 10.1016/j.ijforecast.2018.03.005.

4. KE, J. ir kt. Short-Term Forecasting of Passenger Demand under On-Demand Ride Services: A Spatio-Temporal Deep Learning Approach. *Transportation Research Part C: Emerging Technologies*. 2017, 85, 591–608. Prieiga per: doi: 10.1016/j.trc.2017.10.016.

5. ZENG, A. ir kt. Are Transformers Effective for Time Series Forecasting? [interaktyvus]. .arXiv, 2022arXiv:2205.13504 [cs]. . Prieiga per: http://arxiv.org/abs/2205.13504

6. WEN, Q. ir kt. Transformers in Time Series: A Survey. [interaktyvus]. .arXiv, 2023arXiv:2202.07125 [cs, eess, stat]. . Prieiga per: http://arxiv.org/abs/2202.07125

7. KHAIDEM, L. ir kt. Predicting the direction of stock market prices using random forest. [interaktyvus]. .arXiv, 2016arXiv:1605.00003 [cs]. . Prieiga per: http://arxiv.org/abs/1605.00003

8. XIE, Y. ir kt. Explanation of Machine-Learning Solutions in Air-Traffic Management. *Aerospace*. 2021, 8(8), 224. Prieiga per: doi: 10.3390/aerospace8080224.

9. CHAKRABORTY, D. - ELZARKA, H. Advanced machine learning techniques for building performance simulation: a comparative analysis. *Journal of Building Performance Simulation*. 2019, 12(2), 193–207. Prieiga per: doi: 10.1080/19401493.2018.1498538.

10. LIU, L. ir kt. Combining Partial Least Squares and the Gradient-Boosting Method for Soil Property Retrieval Using Visible Near-Infrared Shortwave Infrared Spectra. *Remote Sensing*. 2017, 9(12), 1299. Prieiga per: doi: 10.3390/rs9121299.

11. WANG, X. - KADIOGLU, S. Dichotomic Pattern Mining with Applications to Intent Prediction from Semi-Structured Clickstream Datasets. [interaktyvus]. .arXiv, 2022arXiv:2201.09178 [cs]. . Prieiga per: http://arxiv.org/abs/2201.09178

12. CHEN, Y. ir kt. Probabilistic Forecasting with Temporal Convolutional Neural Network. [interaktyvus]. .arXiv, 2020arXiv:1906.04397 [cs, stat]. . Prieiga per: http://arxiv.org/abs/1906.04397

13. TUGAY, R. - OGUDUCU, S.G. Demand Prediction Using Machine Learning Methods and Stacked Generalization. [interaktyvus]. .arXiv, 2022arXiv:2009.09756 [cs, stat]. . Prieiga per: http://arxiv.org/abs/2009.09756

14. GU, J. ir kt. Recent Advances in Convolutional Neural Networks. [interaktyvus]. .arXiv, 2017arXiv:1512.07108 [cs]. . Prieiga per: http://arxiv.org/abs/1512.07108

15. HATAMI, N. ir kt. Classification of Time-Series Images Using Deep Convolutional Neural Networks. [interaktyvus]. .arXiv, 2017arXiv:1710.00886 [cs]. . Prieiga per: http://arxiv.org/abs/1710.00886

16. KASHIPAREKH, K. ir kt. ConvTimeNet: A Pre-trained Deep Convolutional Neural Network for Time Series Classification. [interaktyvus]. .arXiv, 2019arXiv:1904.12546 [cs, stat]. . Prieiga per: http://arxiv.org/abs/1904.12546

17. MA, Q. ir kt. A Survey on Time-Series Pre-Trained Models. [interaktyvus]. .arXiv, 2023arXiv:2305.10716 [cs]. . Prieiga per: http://arxiv.org/abs/2305.10716

18. CHUNG, J. ir kt. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. [interaktyvus]. .arXiv, 2014arXiv:1412.3555 [cs]. . Prieiga per: http://arxiv.org/abs/1412.3555

19. GREFF, K. ir kt. LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*. 2017, 28(10), 2222–2232. Prieiga per: doi: 10.1109/TNNLS.2016.2582924.

20. YU, J. A New Way of Airline Traffic Prediction Based on GCN-LSTM. *Frontiers in Neurorobotics* [interaktyvus]. 2021, 15. Prieiga per: https://www.frontiersin.org/articles/10.3389/fnbot.2021.661037

21. QIXIU CHENG ir kt. Analysis and Forecasting of the Day-to-day Travel Demand Variations for Large- scale Transportation Networks: A Deep Learning Approach. [interaktyvus]. 2016. Prieiga per: http://rgdoi.net/10.13140/RG.2.2.12753.53604

22. SUTSKEVER, I. ir kt. Sequence to Sequence Learning with Neural Networks. [interaktyvus]. .arXiv, 2014arXiv:1409.3215 [cs]. . Prieiga per: http://arxiv.org/abs/1409.3215

23. ATLIOĞLU, M.C. ir kt. Supervised Learning Approaches to Flight Delay Prediction. *Sakarya University Journal of Science*. 2020, 24(6), 1223–1231. Prieiga per: doi: 10.16984/saufenbilder.710107.

24. YAN, Z. ir kt. A Deep Learning Approach for Short-Term Airport Traffic Flow Prediction. *Aerospace*. 2022, 9(1), 11. Prieiga per: doi: 10.3390/aerospace9010011.

25. MARCOS, R. ir kt. A Machine Learning Approach to Air Traffic Route Choice Modelling. . . Prieiga per: doi: https://arxiv.org/pdf/1802.06588.

26. DELAHAYE, D. ir kt. Air Traffic Flow Representation and Prediction using Transformer in Flow-centric Airspace. [interaktyvus]. . Prieiga per: https://enac.hal.science/hal-03907364/document

27. HEFFAR, M. ir kt. Prediction of Flight Departure and Arrival Routes with Gradient Boosted Decision Trees. [interaktyvus]. 2021. Prieiga per: https://www.researchgate.net/profile/Ramon-Dalmau-Codina/publication/356782642_Prediction_of_Flight_Departure_and_Arrival_Routes_with_Gradient_Boosted_Decision_Trees/links/61ab9e8250e22929cd47e62a/Prediction-of-Flight-Departure-and-Arrival-Routes-with-Gradient-Boosted-Decision-Trees.pdf\

28. GUNTER, U. - ZEKAN, B. Forecasting air passenger numbers with a GVAR model. *Annals of Tourism Research*. 2021, 89, 103252. Prieiga per: doi: 10.1016/j.annals.2021.103252.

29. Features — LightGBM 4.3.0.99 documentation. [interaktyvus]. [žiūrėta 2024-05-11]. Prieiga per: https://lightgbm.readthedocs.io/en/latest/Features.html#optimal-split-for-categorical-features

30. CERDA, P. - VAROQUAUX, G. Encoding high-cardinality string categorical variables. *IEEE Transactions on Knowledge and Data Engineering*. 2022, 34(3), 1164–1176. Prieiga per: doi: 10.1109/TKDE.2020.2992529.

31. GUO, C. - BERKHAHN, F. Entity Embeddings of Categorical Variables. [interaktyvus]. .arXiv, 2016arXiv:1604.06737 [cs]. . Prieiga per: http://arxiv.org/abs/1604.06737

32. MAHAJAN, T. ir kt. An Experimental Assessment of Treatments for Cyclical Data. [interaktyvus]. . Prieiga per: https://scholarworks.calstate.edu/downloads/pv63g5147

Priedai

1 priedas. Priedo pavadinimas

Kol kas nieko.