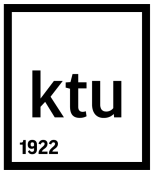


KAUNO TECHNOLOGIJOS UNIVERSITETAS
ELEKTROS IR ELEKTRONIKOS FAKULTETAS

Mantas Butkus

Įmonės mokumo problemų identifikavimo sistemos tyrimas

Baigiamasis magistro projektas

Vadovas
Doc. dr. Vygandas Vaitkus

KAUNAS, 2018

KAUNO TECHNOLOGIJOS UNIVERSITETAS
ELEKTROS IR ELEKTRONIKOS FAKULTETAS
AUTOMATIKOS KATEDRA

Įmonės mokumo problemų identifikavimo sistemos tyrimas

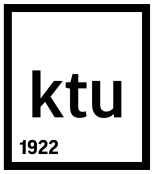
Baigiamasis magistro projektas
Valdymo technologijos (kodas 621H66001)

Vadovas
Doc. dr. Vygandas Vaitkus
2018-05-27

Recenzentas
Prof. Rimvydas Simutis

Projektą atliko
Mantas Butkus
2018-05-27

KAUNAS, 2018



KAUNO TECHNOLOGIJOS UNIVERSITETAS
ELEKTROS IR ELEKTRONIKOS FAKULTETAS
Mantas Butkus
Valdymo technologijos, kodas 621H66001

Baigiamojo projekto „Įmonės mokumo problemų identifikavimo sistemos tyrimas“

AKADEMINIO SĄŽININGUMO DEKLARACIJA

2018 m. gegužės 27 d.
Kaunas

Patvirtinu, kad mano, Manto Butkaus, baigiamasis projektas tema „Įmonės mokumo problemų identifikavimo sistemos tyrimas“ yra parašytas visiškai savarankiškai, o visi pateikti duomenys ar tyrimų rezultatai yra teisingi ir gauti sąžiningai. Šiame darbe nei viena dalis nėra plagijuota nuo jokių spausdintinių ar internetinių šaltinių, visos kitų šaltinių tiesioginės ir netiesioginės citatos nurodytos literatūros nuorodose. Įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs.

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

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

Butkus, Mantas. Įmonės mokumo problemų identifikavimo sistemos tyrimas. *Valdymo sistemų magistro* baigiamasis projektas / vadovas doc. dr. Vygandas Vaitkus; Kauno technologijos universitetas, Elektros ir elektronikos fakultetas, Automatikos katedra.

Mokslo kryptis ir sritis: Elektros ir elektronikos inžinerija, Technologiniai mokslai

Reikšminiai žodžiai: *sąskaita faktūra, mašininis mokymas, klasifikavimas*

Kaunas, 2018. 48 psl.

SANTRAUKA

Sąskaitų faktūrų apmokėjimas laiku yra svarbus veiksnys, darantis įtaką įmonės mokumui. Bankas, suteikęs paskolą ar kreditą, neanalizuoja savo kliento sąskaitų apmokėjimo, kol kompanija neatsiduria arti nemokumo ribos, tačiau tada įmonei išvengti bankroto darosi sudėtinga. Siekiant identifikuoti įmonės mokumo problemas kuriamos sistemos, kurios perspėtų banką, jog jo kliento mokumas artėja prie pavojingos ribos. Bakalauro baigiamojo projekto metu buvo sukurta sąskaitų faktūrų apmokėjimo prognozavimo sistema, kuri įmonių mokumui prognozuoti naudoja tik sąskaitų faktūrų duomenis. **Šio projekto tikslas –** atlikti papildomus tyrimus taikant įvairius ne tik regresijos, bet ir klasifikavimo modelius su tikslu pagerinti mokumo problemų turinčių įmonių aptikimo sistemos kokybę. Šiam tikslui įgyvendinti buvo iškelti uždaviniai:

1. ištirti mokslines publikacijas, kuriose įmonės mokumas prognozuojamas panaudojant tik mokėjimų informaciją;
2. sukurti bei ištirti įmonės mokumo prognozavimo sistemą panaudojant klasifikavimo modelius;
3. sukurti adaptyvią mokumo nustatymo sistemą;
4. atlikti neapmokėtų sąskaitų faktūrų aptikimo sistemos kokybės vertinimą.

Sukurtos sistemos veikimas ištirtas panaudojant sprendimo medžių kolektyvo, atraminių vektorių mašinos, Bajeso klasifikatoriaus bei gilaus mokymosi neuroninių tinklų modelius. Nustatyta, jog sėkmingam sistemos veikimui pakankamas 5–6 požymių sąrašas, o optimalus slenkstis (angl. *treshhold*) atrenkant reikšmingus požymius yra 0,35. Geriausi rezultatai pasiekti naudojant sprendimo medžių kolektyvo bei atraminių vektorių mašinos modelius, tuo tarpu gilaus mokymo niauroninis tinklas gali būti taikomas įmonės mokumui prognozuoti. Darbo rezultatai publikuotas Kauno technologijos universitete rengiamoje E^2TA konferencijoje.

Butkus, Mantas. *Research of company's solvency problem identification system*: *Master's* thesis in *Control systems* / supervisor doc. dr. Vygandas Vaitkus. Kaunas University of Technology, Faculty of Electrical and Electronics Engineering, department of Automation.

Research area and field: Electrical and Electronics Engineering, Technological Sciences

Key words: invoice payment, machine learning, classification

Kaunas, 2018. 48 p.

SUMMARY

On time payment of invoices is an important factor that influences a company's solvency. Banks usually do not follow their customer payments until the customer pays his invoices. When the bank notices that his customer invoices are not paid usually it is too late and the company goes bankrupt. To avoid this it is important to identify companies that have solvency problems. In order to identify these companies company solvency prediction models are made. These systems alert the bank, that a company may have solvency problems in the near future. The aim of this work is to research a company's solvency problem identification system using not only regression but also classification models. Following tasks were created:

1. research company solvency detection systems that use only invoice information in other publications;
2. create a classification model for the current system;
3. create an adaptive company solvency problem detection system;
4. evaluate the quality of the system.

Decision tree ensemble, support vector machine, Bayes and deep learning neural network models were used and analysed. It was determined that 5 to 6 features are sufficient for a successful performance of the system. The optimal threshold for the importance of the feature is 0.35. The best forecasting results were obtained using the decision tree ensemble and support vector machine methods, however the deep learning neural network model may also be used for invoice payment prediction. This project was also published at the E^2TA conference in Kaunas University of Technology.

TURINYS

ĮVADAS

Bankas, suteikęs klientui (pvz. įmonei) paskolą, neanalizuoja savo kliento mokumo, jei klientas laiku moka įnašus. Tik tam tikrą laiką neapmokėjus sąskaitų, bankas pradeda domėtis savo kliento finansine būkle, tačiau dažnai kliento mokumui jau yra iškilusi grėsmė. Išnaudojami gauti papildomi kreditai. Klientui gresia bankrotas, o paskolą išdavusiam bankui – finansiniai nuostoliai prarandant ne tik sukeiktą paskolą, bet ir kreditus. Siekiant išspręsti šią problemą pastaruoju metu mokslininkai ir finansų analitikai vis dažniau bando išsiaiškinti įmonių bankroto galimybes, t.y. atlieka jų prognozavimą [1, 2, 3, 4]. Tačiau reikia paminėti, kad didžioji dalis tokių tyrimų remiasi kompanijos finansiniais rodikliais, kurie gaunami kas tris mėnesius ar rečiau. Naudojant šias sistemas įmonės mokumo problemos identifikuojamos pavėluotai. Nagrinėjant įmonės sąskaitų faktūrų informaciją, įmonės mokumui identifikuoti reikalingi duomenys pasiekiami lengvai, jei banko paskolos suteikimo sąlygose nurodytą, jog klientas savo sąskaitas apmokės per banko nurodytą įmonę, tačiau tik labai mažai tyrimų atliekama naudojant įmonės mokėjimų duomenis [5, 6, 7, 8].

Bakalauro baigiamojo projekto metu buvo sukurta sąskaitų faktūrų apmokėjimo prognozavimo sistema, kuri įmonių mokumui prognozuoti naudoja tik sąskaitų faktūrų duomenis [9]. **Šio projekto tikslas** – atlikti papildomus tyrimus taikant įvairius ne tik regresijos, bet ir klasifikavimo modelius su tikslu pagerinti mokumo problemų turinčių įmonių aptikimo sistemos kokybę. Šiam tikslui įgyvendinti buvo iškelti uždaviniai:

1. išanalizuoti teorinę medžiagą, kuriose įmonės mokumas prognozuojamas panaudojant tik mokėjimų informaciją;
2. sukurti bei ištirti įmonės mokumo prognozavimo sistemą panaudojant klasifikavimo modelius;
3. sukurti adaptyvią mokumo nustatymo sistemą;
4. atlikti neapmokėtų sąskaitų faktūrų aptikimo sistemos kokybės vertinimą.

Analizuojama sistema remiasi regresijos modeliu bei naudoja apibrėžtą kiekį požymių, kuriems apskaičiuoti reikia daug laiko, tačiau nėra aišku, ar visi požymiai tikrai reikalingi, kad sistema sėkmingai veiktų. Įmonės mokumas prognozuojamas tik 30-čiai dienų ir nėra galimybės keisti šį laiko tarpą. Siekiant pagerinti ir išanalizuoti sistemos veikimą atliekamas šios sistemos tyrimas panaudojant ne tik regresijos, bet ir klasifikavimo bei gilaus mokymosi metodus iškeltai problemai spręsti.

1. LITERATŪROS APŽVALGA

Nagrinėjamai darbo temai suprasti reikalingos ne tik procesų valdymo ir optimizavimo, tačiau ir statistinės žinios bei verslo procesų supratimas. Taigi kitame poskyryje aptariamos svarbiausios magistro baigiamajame projekte naudojamos sąvokos, susijusios su finansais ir verslu.

1.1. Sąvokos

Sąskaita faktūra (SF) – apskaitos dokumentas, kuriuo įforminamas prekių tiekimas ar paslaugų teikimas. Sąskaitą faktūrą vartotojui išrašo prekes pateikęs ar paslaugas suteikęs asmuo, reikalaudamas sumokėti sąskaitoje faktūroje nurodytą sumą [10].

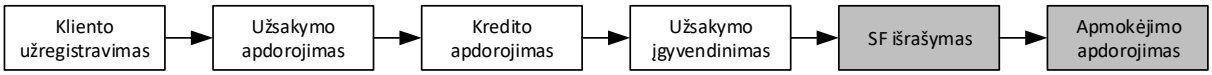
SF apmokėjimo terminas – data, iki kurios turi būti apmokėta SF. Šis terminas gali apimti skirtingą laikotarpį, pvz., 14, 30 ar 90 dienų.

Neapmokėta SF (NSF) – tai sąskaita faktūra, kuri nebuvo laiku apmokėta.

Apmokėta SF (ASF) – tai sąskaita faktūra, kuri buvo apmokėta per SF nustatytus apmokėjimo terminus.

1.2. Kredito rizikos sąsaja su neapmokėtomis sąskaitomis faktūromis

Atliekant kredito suteikimo rizikos analizę, klientai skirstomi į skirtingas grupes, remiantis tam tikromis kliento ypatybėmis – sąskaitos balansu, mokėjimų istorija ir kita finansinėse ataskaitose pateikiama informacija [1]. Pavėluotai apmokėti ar neapmokėti mokėjimai yra glaudžiai susiję. Analizuojant neapmokėtas SF, klientai skirstomi į grupes, priklausomai nuo to, ar SF buvo apmokėta per nustatytus terminus. Tipinės SF išrašymo ir apmokėjimo procedūros schema pateikiama 1.1 paveiksle.

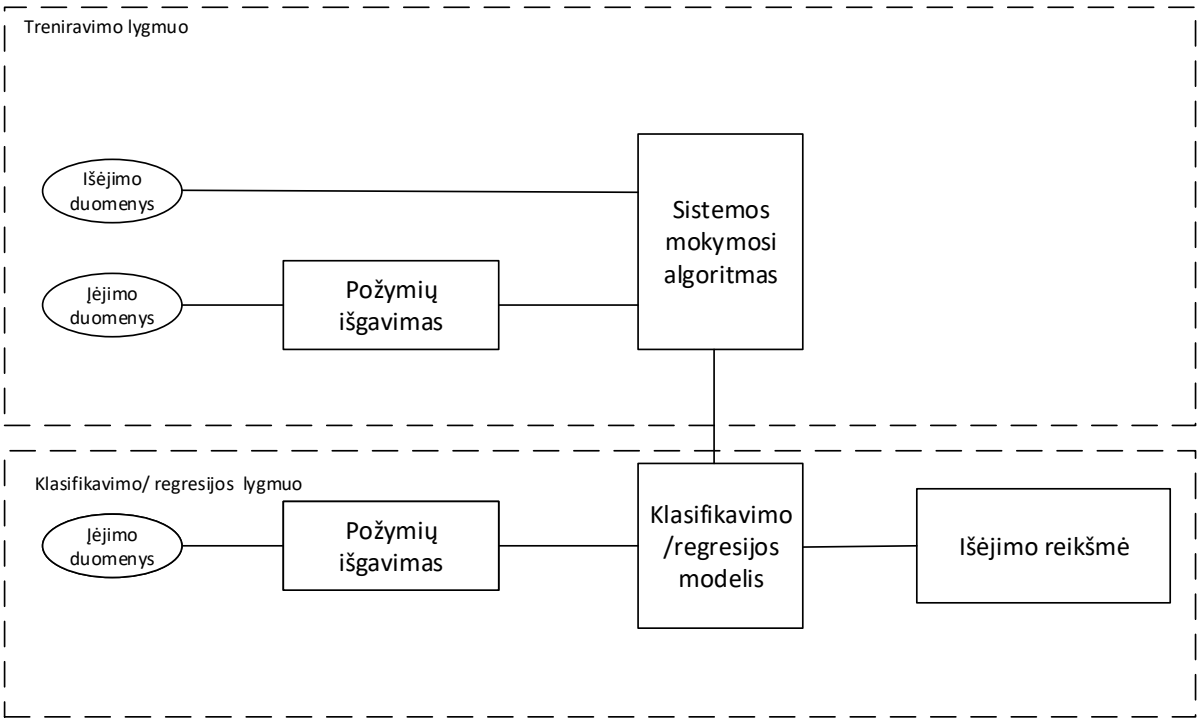


1.1 pav. Sąskaitos faktūros apmokėjimo procedūra. Adaptuota iš [8]

Šiame projekte nagrinėjama tik pilkai pažymėta sritis, kadangi orientuojamasi į SF mokėjimų informaciją. SF apmokėjimas užima svarbų vaidmenį kasdieniniame versle, nes NSF gali sukelti įmonei likvidumo problemų. Gebėjimas prognozuoti pinigų srautus yra ypač svarbus norint išlaikyti finansinį stabilumą. Remiantis Europos mokėjimų ataskaita 2016 (angl. *European Payment report 2016*) [11], 40 % respondentų teigia, jog NSF trukdo įmonės plėtimuisi. 33 % apklaustų įmonių mano, jog NSF kelia didelę grėsmę įmonės išlikimui ilgalaikėje perspektyvoje. Remiantis apklaustaisiais, įmonė praranda apie 2,2 % pelno dėl NSF. Šiai problemai spręsti bei siekiant sumažinti patiriamus nuostolius reikalinga sistema, gebanti prognozuoti SF neapmokėjimą. Tokiu būdu būtų galima imtis veiksmų, kol dar ne per vėlu ir įmonė nebankrutavo bei bankas nepatyrė nuostolių.

1.3. Neapmokėtų SF prognozavimas

Įmonės bankroto prognozavimas buvo nagrinėjamas dar 60-ųjų pabaigoje, kai prognozavimui buvo panaudojami viešai prieinami duomenys ir statistiniai klasifikavimo metodai. Vienas iš pirmųjų darbų, siekiančių atlikti modernią statistinę bankroto prognozavimo analizę, buvo parašytas Tamari [2]. Nuo 2000 -ųjų pastebimas stabilus išleistų publikacijų, susijusių su kreditavimo rizika, augimas. Tai galėjo atsitikti dėl didelio susidomėjimo kreditavimo rizikos modeliavimu ir naujų duomenų surinkimo metodų atsiradimo bei kredito rinkų augimo [8]. Publikacijų, kuriose būtų aprašytas SF neapmokėjimo prognozavimas panaudojant tik mokėjimų duomenis, išleista nedaug. Nagrinėtose publikacijose [6, 7, 8] SF apmokėjimą prognozuojanti sistema kuriama pasitelkiant 1.2 paveiksle pavaizduotą modelį.



1.2 pav. Modelio sudarymo struktūra. Adaptuota iš [7]

Sistemą galima suskaidyti į treniravimo ir klasifikavimo arba regresijos lygmenis. Pirmajame lygmenyje duomenys suskaidomi į treniravimo ir testavimo duomenis. Panaudojant požymių išgavimo algoritmą paruošiami modelio įėjimo duomenys, kurie naudojami kartu su atitinkamais išėjimo duomenimis pasirinktam modeliui apmokyti. Atlikus sistemos apmokymą ir pasiekus norimą tikslumą, atliekamas modelio testavimas, kur modelio įėjimai yra sistemos treniravimui nenaudoti duomenys. Gautos modelio išėjimo reikšmės lyginamos su tikrosiomis, įvertinant modelio tikslumą.

1.4. Duomenų apdorojimas

Priklausomai nuo turimo kiekio, duomenys atliktuose tyrimuose pertvarkomi, atliekant statistinę analizę, nagrinėjant, kiek SF turima ir kiek iš jų yra neapmokėta. Zengo [8] kuriamame

modelyje naudojami 4 įmonių SF duomenys, kurių apžvalga pateikiama 1.1 lentelėje.

1.1 lentelė. SF kiekiai įmonėse. Adaptuota iš [8]

Įmonė	SF kiekis
A	40908
B	109589
C	22701
D	8474

Zengo tyrime nepateikiama, kiek iš šių SF nebuvo apmokėta, taigi negalima spręsti apie duomenų subalansuotumą. P. Hu [7] ir W. Hu [6] kuriamose sistemose buvo panaudojami Fortune 500 įmonių grupės duomenys. Duomenų apžvalga pateikiama 1.2 lentelėje:

1.2 lentelė. SF kiekiai įmonėse

	SF kiekis	Apmokėtos SF	Neapmokėtos SF
P. Hu duomenys [7]	72464	19565,(27%)	52899,(73%)
W. Hu duomenys [6]	4446045	3609123 (81,2 %)	836922 (18,8 %)

Kaip matyti iš lentelėje pateiktų duomenų, P. Hu atliekamo tyrimo duomenyse vyrauja NSF, tuo tarpu Hu, W. naudojami duomenys pasižymi nedideliu NSF kiekiu, taigi jų aptikimas tampa sudėtingesnis.

1.5. Požymių sudarymas

Atlikus duomenų statistinę analizę formuojami modelio įėjimo duomenys. Modelio įėjimams sudaromas duomenų masyvas – požymių sąrašas. Šiame sąraše galima išskirti SF ir kliento lygmenis [8]. SF lygmenį sudaro tik SF pateikti duomenys, šiame lygmenyje galima išskirti tokius požymius [7, 6]:

1. SF suma;
2. SF išrašymo data;
3. data, iki kurios turi būti apmokėta SF;
4. apmokėjimo periodo trukmė.

Šie požymiai sudaromi iš atskiros SF, nenagrinėjant kliento istorinių duomenų. Antrajame, kliento, lygmenyje, požymiai parengiami naudojant istorinius SF duomenis kiekvienam klientui. Zengo tyrime nustatyta, jog šių požymių įtraukimas į modelį žymiai padidina modelio prognozavimo tikslumą [8].

1.3 lentelė. Prognozavimo rezultatai, panaudojant SF istorinius duomenis. Adaptuota iš [8]

Kompanija	Požymių kategorija	Tikslumas, %
A	Sąskaita faktūra	68,24
	Sąskaita faktūra + istorija	81,38
B	Sąskaita faktūra	84,72
	Sąskaita faktūra + istorija	88,28
C	Sąskaita faktūra	49,68
	Sąskaita faktūra + istorija	66,46
D	Sąskaita faktūra	58,79
	Sąskaita faktūra + istorija	70,87

Kaip matyti 1.3 lentelėje, prognozavimo rezultatai žymiai didesni, naudojant SF istorinius duomenis. Visuose išnagrinėtuose tyrimuose [5, 6, 7, 8] įėjimo duomenys formuojami panaudojant požymių sąrašą, kurį sudaro tiek požymiai, sudaryti iš SF duomenų, tiek iš jos istorinių duomenų, taigi panaudojami abu SF lygmenys. Papildomai sugeneruojami ir kalendoriniai požymiai, kurie panaudojami Hu [6, 7] prognozavimo sistemose.

1.5.1. Mėnesio pabaigos indikatorius

Mėnesio pabaigos indikatorius – šis požymis nurodo, ar SF yra išrašyta mėnesio pabaigoje. Mėnesio pabaigos indikatorius gali įgauti 0 arba 1 reikšmę. Jam priskiriama 1 reikšmė, jei SF buvo išrašyta per tris dienas iki mėnesio pabaigos. Šis požymis sukuriamas remiantis prielaida, jog mėnesio pabaigoje gali būti sunkiau apmokėti SF, taigi jos apmokėjimas gali užsitęsti.

1.5.2. Antros mėnesio pusės indikatorius

Antros mėnesio pusės indikatorius parodo, kurioje mėnesio pusėje SF buvo išrašyta. Šis indikatorius lygus 1, jei SF yra išrašoma vėliau nei penkioliktą mėnesio dieną, bet ne vėliau nei prieš tris paskutines mėnesio dienas. Šis požymis sudaromas remiantis prielaida, jog antroje mėnesio pusėje įmonė finansiškai stabilesnė nei mėnesio pradžioje, kadangi neseniai buvo mokami atlyginimai darbuotojams ir mėnesio pabaigoje, kai tenka mokėti tiek atlyginimus, tiek apmokėti savas sąskaitas.

Aptarti požymiais kombinuojami į vieną sąrašą. Tyrimuose naudojami požymių sąrašai pateikiami 1.4-oje lentelėje.

1.4 lentelė. Tyrimuose naudojami požymių sąrašai

Zengo tyrimas [8]	W. Hu tyrimas [6]	P. Hu tyrimas [7]
SF suma	SF suma	SF suma
SF apmokėjimo terminas	mėn. pabaigos indikatorius	mėn. pabaigos indikatorius
Kategorija	mėn. antros pusės indikatorius	mėn. antros pusės indikatorius
ASF kiekis	SF kiekis	SF kiekis
NSF kiekis	NSF kiekis	NSF kiekis
ASF suma	bendra SF suma	bendra SF suma
NSF suma	vidutinė SF suma	bendra NSF suma
santykis tarp 8 ir 7	bendra NSF suma	vidutinė SF suma
vidutinis ASF vėlavimas	vidutinė NSF suma	vidutinė NSF suma
	vėlavimo santykis	vidutinis SF vėlavimas
	sumų santykis	vidutinis NSF vėlavimas
	vidutinis apmokėjimo terminas	santykis tarp 5 ir 4
	vidutinis vėlavimas	santykis tarp 8 ir 7
		vidutinis apmokėjimo terminas
		Sektorius
		Įmonės pavadinimas

Kaip matyti lentelėje, visuose tyrimuose naudojami panašūs požymiai. Kombinuojami tiek SF duomenys (SF suma, SF sritis), tiek kliento istoriniai SF duomenys, skaičiuojant statistinius parametrus (SF kiekiai, sumos bei vidurkiai ar santykiai tarp šių dydžių) bei aptarti kalendoriniai indikatoriai.

1.6. Modelio parinkimas

Paruošus sistemos įėjimo požymių sąrašą, parenkamas sistemoje naudojamas modelis. Išanalizuotoje literatūroje sprendžiamas klasifikavimo uždavinys [5, 6, 7, 8]. Čia SF skirstomos į keturias klases pagal jų apmokėjimą:

1. laiku apmokėtos SF;
2. SF, kurių apmokėjimas vėluos iki 30 dienų;
3. SF, kurių apmokėjimas vėluos nuo 30 iki 90 dienų;
4. SF, kurios nebus apmokėtos po 90 dienų.

SF apmokėjimui prognozuoti buvo panaudoti aštuoni skirtingi modeliai:

1. sprendimų medžiai;
2. atsitiktiniai miškai;
3. adaptinis bustingas (angl. *adaptive boosting*);
4. logistinė regresija;
5. atraminių vektorių mašina (angl. *Support vector mashine*);
6. artimiausio K kaimyno (angl. *K nearest neighbour*) modelis;
7. neuroniniai tinklai;

8. atsitiktiniai miškai su skirtingais svoriais (angl. *weighted random forests*).

Pirmasis nagrinėjamas modelis yra sprendimų medis. Tai modelis, kuris panašus į medį, kurio kiekviena šaka yra klasifikavimo klausimas. Medis optimizuojamas parenkant optimalų šakų skaičių [12]. Šiuo atveju pagrindiniai požymiai, kuriuos naudojo sprendimų medis, buvo kliento vėlavimo santykis, vidutinis vėlavimas ir kliento bendras sąskaitų faktūrų skaičius. Šis modelis pasiekė 86,1 % tikslumą. Atsitiktiniai miškai – tai klasifikavimo metodas, kuris remiasi sprendimų medžiais. Sudaroma keletas sprendimo medžių, kurie „nubalsuoja“ ar sąskaita faktūra bus apmokėta, ar ne ir priskiria atitinkamai grupei [13]. Svarbiausi požymiai šiam modeliui buvo kliento vėlavimo santykis ir vidutinis kliento apmokėjimo vėlavimas dienomis. Šis modelis pasiekė 89,2 % tikslumą ir buvo tiksliausias. Adaptinio bustinimo metodas paremtas tuo, jog naudojant kelių tipinių modelių sumą – sudaroma nauja, stipresnė prognozė [14]. Šis metodas pasiekė 86,3 % tikslumą. Logistinė regresija – tai toks modelis, kur vienas ar keletas nepriklausomų kintamųjų daro įtaką vienam dvireikšmiam kintamajam [13]. Naudojant šį metodą didžiausią reikšmę turėjo mėnesio pabaigos ir vidurio indikatorių požymiai, taip pat kliento bendras sąskaitų faktūrų skaičius, vėlavusių sąskaitų faktūrų skaičius, vidutinis sąskaitų faktūrų apmokėjimo vėlavimas dienomis ir bendrų ir vėlavusių sąskaitų faktūrų santykis. Šis metodas pasiekė 86,4 % tikslumą. Atraminių vektorių mašinos modelis Hu sistemoje pasiekė 86,9% tikslumą. Prognozavimo rezultatų apžvalga pateikiama 1.5 lentelėje.

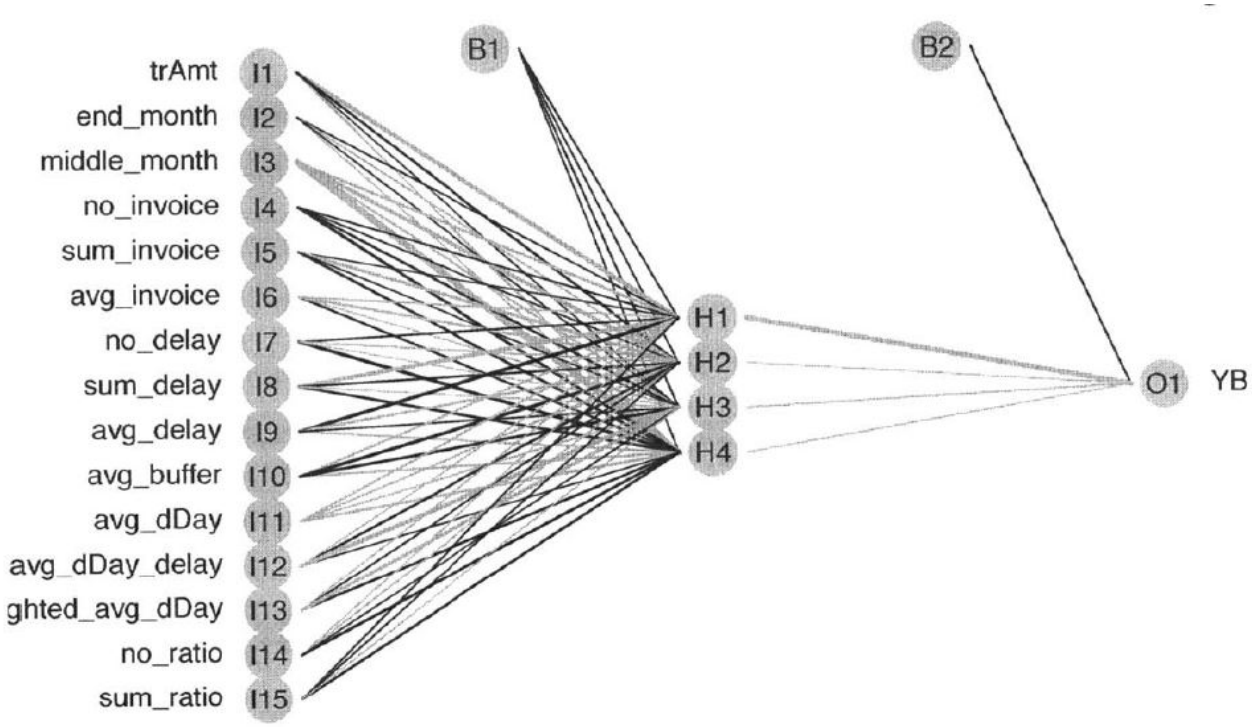
1.5 lentelė. Prognozavimo rezultatai. Adaptuota iš [7]

Modelis	Tikslumas, %
Sprendimų medis	86,1
Atsitiktiniai miškai	89,2
Adaptinio bustinimo metodas	86,3
Logistinė regresija	86,4
Atraminių vektorių mašina	86,9

Kaip matyti iš lentelėje pateiktų duomenų tiksliausiai (81,6 %) prognozavo atsitiktinių miškų modelis.

Hu W. sukurtoje sistemoje taip pat naudotas dirbtinio neuroninio tinklo modelis. Dirbtiniai neuroniniai tinklai laikomi informacijos apdorojimo metodu, kuris paremtas smegenyse vykstančių procesų imitavimu [15]. Dirbtinis neuroninis tinklas yra matematinių modelių rinkinys, kurio pagalba bandoma imituoti gyvų organizmų gebėjimą mokytis, prisitaikyti bei adaptuotis. Dirbtiniai neuroniniai tinklai sudaryti iš daugelio sujungtų elementarių skaičiuojamųjų elementų – neuronų. Šie elementai paremti biologiniais neuronais ir jungiasi vieni su kitais sudarydami įvairaus stiprumo jungtis, kurios yra analogiškos biologinių neuronų jungtims. Biologinėse sistemose mokymosi metu smegenyse keičiasi jungčių, siejančių neuronus, stiprumas. Analogiškai atliekamas ir dirbtinio neuroninio tinklo mokymas – jungtys keičiamos, kol pasiekiami tenkinantys rezultatai. Neuroninių tinklų mokymui naudojami įėjimo – išėjimo duomenų pavyzdžiai, pagal kurios specialių algoritmų pagalba mokymo metu iteratyviai keičiami jungčių stiprumo koeficientai arba svoriai. Informacija, reikalinga konkretaus uždavinio sprendimui, yra sukaupiama šių svorių

reikšmėse [16]. Naudotos sistemos neuroninis tinklas pateikiamas 1.3 – iame paveiksle.



1.3 pav. Tyrime naudotas neuroninis tinklas [6]

Kaip matyti paveiksle, neuroninį tinklą sudaro 15 įėjimų neuronų, 4 neuronai paslėptame sluoksnyje ir vienas neuronas išėjimo sluoksnyje. Hu, W. tyrimo metu gauti prognozavimo rezultatai pateikiami 1.6-oje lentelėje:

1.6 lentelė. Prognozavimo rezultatai. Adaptuota iš [6]

Modelis	Tikslumas, %
Atsitiktiniai miškai	75
Atsitiktiniai miškai su skirtingais svoriais	78,29

Tiek aprašyto neuroninio tinklo, tiek kitų anksčiau paminėtų modelių prognozavimo rezultatai W. Hu daktaro disertacijoje nepateikiami. Iš 1.6 lentelėje pateiktų duomenų matyti, jog atsitiktinių miškų su skirtingais svoriais metodas pagerino savo pirmtako rezultatus.

2. SISTEMOJE NAUDOJAMŲ DUOMENŲ APŽVALGA

Analizuojamos sistemos tyrimas buvo atliktas panaudojant realius įmonių istorinius sąskaitų faktūrų mokėjimų duomenis, kurie buvo gauti iš mokslinio tiriamojo projekto užsakovo *CSV* formatu. Duomenų rinkinį sudaro 1181538 SF informacija. Duomenys buvo analizuojami pasitelkiant MATLAB programavimo aplinką. Gautame CSV faile pateikti duomenys apima penkerių metų laikotarpį nuo 2010-01-01 iki 2015-01-01. Kelių SF duomenų masyvo eilučių pavyzdys pateikiamas 2.1 lentelėje.

2.1 lentelė. Duomenų paketo fragmentas

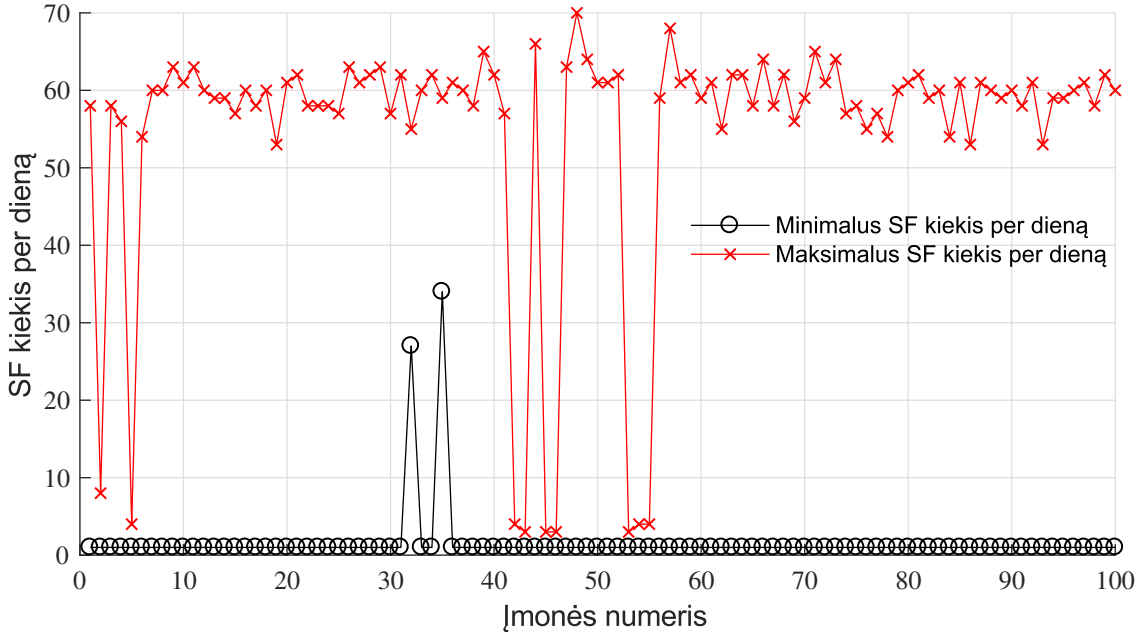
FINA18	15513	25-Oct-2012	01-Dec-2012	100	NaN	NaN	1	90
RETA1	-89904	02-Jan-2010	10-Jan-2010	52,01	15-Feb-2010	47,99	2	14

Pateikiamu atveju, SF sudaro įmonės identifikavimo kintamasis, kuris parodo, kuriam sektoriui ši įmonė priklauso. Pirmoje eilutėje pateikiami finansų sektoriaus 18-os įmonės duomenys, antroje – mažmeninės prekybos sektoriaus pirmos įmonės duomenys. Antrame stulpelyje pateikiama sąskaitos faktūros suma eurais. Ji turi būti sumokėta iki nustatyto termino pabaigos. Nagrinėjamuose duomenyse skiriamos dvi SF rūšys – gauta ir išsiųsta. Atliekant prognozavimą, atsižvelgiama tik į neigiamas reikšmes turinčias SF, kurios atitinka gautąsias. Vertinamos tik šios SF, nes mus domina įmonės mokumas, tai yra, ar įmonė apmokės gautą SF, o ne jos klientų mokumas. Toliau pateikiama sąskaitos išrašymo data ir data, kada ji buvo apmokėta. Jei visa SF apmokama iškart, kitame stulpelyje įrašoma 100 % reikšmė. Šis stulpelis parodo, kokia dalis SF sumos buvo apmokėta pirmu kartu. 6 ir 7 stulpeliai nurodo antro mokėjimo informaciją – kada buvo atliktas 2 mokėjimas ir kokia SF dalis buvo apmokėta. Išanalizavus duomenis nustatyta, jog SF apmokėti naudojami mokėjimo terminai yra 14, 30, 60, 90 ir 180 dienų. Šis terminas nurodomas paskutiniame stulpelyje. Taip pat nurodoma, ar sąskaita faktūra gauta iš pirminės ar antrinės įmonės. Pirminė įmonė laikoma svarbesne. Išanalizavus gautus duomenis nustatyta, jog duomenų bazę sudaro 100 įmonių iš aštuonių skirtingų rinkos sektorių sąskaitos faktūros. Jų pasiskirstymas pateikiamas 2.2 lentelėje.

2.2 lentelė. 100 tiriamų įmonių pasiskirstymas sektoriuose

Sektorius	Įmonių skaičius sektoriuje
Mažmeninė prekyba	8
Finansai, draudimas	22
Statyba	1
Transportas, komunalinės paslaugos	9
Paslaugos	20
Gamyba	35
Žemės ūkis, miškininkystė, žuvininkystė	1
Leidyba	4

Dauguma kompanijų priklauso finansų, draudimo, paslaugų ir gamybos sektoriams. Statybos, žemės ūkio, miškininkystės, žuvininkystės sektoriai turi tik po vieną įmonę. Kiekvienos įmonės SF kiekis per dieną svyruoja ir nėra pastovus. Vidutinis SF kiekis per dieną kiekvienoje įmonėje pateikiamas 2.1 paveiksle.



2.1 pav. Maksimalus ir minimalus SF kiekis per dieną

Kaip matyti iš atvaizduotų duomenų, SF kiekis per dieną svyruoja nuo 1 iki 70 vienai įmonei. Taigi per dieną įmonė gauna bent vieną SF, kurią turi apmokėti per nustatytą apmokėjimo terminą.

Parenkant modelį svarbu atsižvelgti, kokią duomenų dalį sudaro NSF. 2.3 lentelėje pateikiami SF kiekiai kiekviename sektoriuje:

2.3 lentelė. ASF ir NSF kiekiai sektoriuose

Sektorius	SF kiekis	NSF kiekis	ASF kiekis	NSF kiekis, %
Mažmeninė prekyba	24628	4747	19881	19,27
Finansai, draudimas	118551	5061	113490	4,26
Statyba	26206	164	26042	0,06
Transportas, komunalinės paslaugos	76653	3975	72678	5,18
Paslaugos	130231	3533	126698	2,71
Gamyba	183951	8282	175669	4,5
Žemės ūkis, miškininkystė, žuvininkystė	5080	535	4545	10,53
Leidyba	38603	559	38044	1,44

Iš lentelės duomenų matyti, jog daugiausia NSF yra mažmeninės prekybos ir žemės ūkio, miškininkystės, žuvininkystės sektoriuose. Kituose sektoriuose NSF kiekis nesiekia 10 %. Daugiausia SF išrašyta gamybos, mažiausiai – žemės ūkio, miškininkystės, žuvininkystės sektoriuje.

3. REGRESIJOS MODELIO APŽVALGA

Sudarant įmonės mokumo problemų identifikavimo modelį galima rinktis tarp literatūros analizėje aptartų klasifikavimo ir regresijos modelių. Bakalauro baigiamojo projekto metu sukurtoje sistemoje buvo naudojamas sprendimo medžio regresijos modelių kolektyvas, kadangi sprendimo medžiai yra vienas iš dažniausiai naudojamų modelių dirbant su nesubalansuotais duomenimis ir anomalijų aptikimu [17].

3.1. Modelio slankiųjų langų sudarymas

Bakalauro baigiamajame projekte sukurtoje sistemoje buvo apibrėžti du slankieji langai, kurie buvo naudojami sudarant modelio įėjimo ir išėjimo duomenis:

1. modelio įėjimo duomenų slankusis langas;
2. modelio išėjimo duomenų slankusis langas.

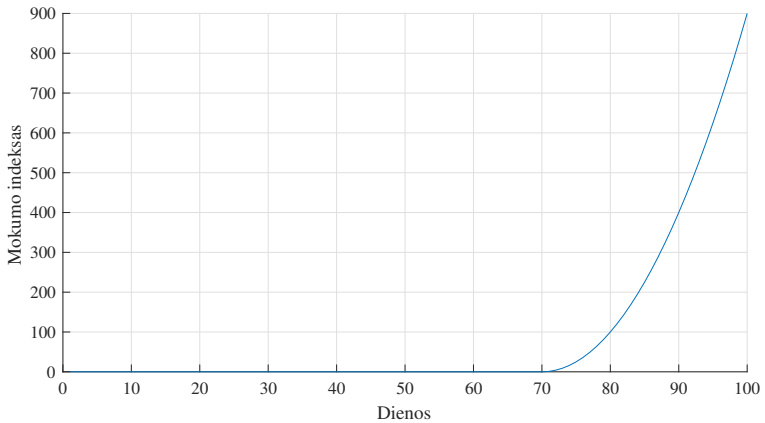
Abu šie langai buvo 30 dienų trukmės ir nurodė, iš kokio laikotarpio buvo skaičiuojami modelio įėjimo bei išėjimo duomenys. Sukurtoje sistemoje šis laiko tarpas yra pastovus ir nekeičiamas, taigi šiame darbe siekiama nustatyti, kokią įtaką mokumo indekso prognozavimui turi šių langų trukmė, todėl šie du langai bus keičiami tarp 30, 60 ir 90 dienų.

3.2. Modelio išėjimo kintamojo formavimas

Kadangi nagrinėjamas regresijos modelis, sistemos išėjimas turėtų būti tolydžiai laike kintantis dydis, identifikuojantis apie kompanijoje atsirandančias mokumo problemas. Tai galėtų atspindėti neapmokėtų SF suma, kiekis arba neapmokėtų ir apmokėtų SF sumų ar kiekių santykis. Analizuojamu atveju buvo pasirinktas neapmokėtų ir apmokėtų SF sumų santykis, kadangi renkantis kiekį, neapmokėta mažos vertės SF būtų tiek pat reikšminga, kaip ir didelės vertės SF. Modelio išėjimo kintamasis aprašomas (3.1) formule:

$$Mokumo\_indeksas = \frac{NSF suma}{ASF suma}. \tag{3.1}$$

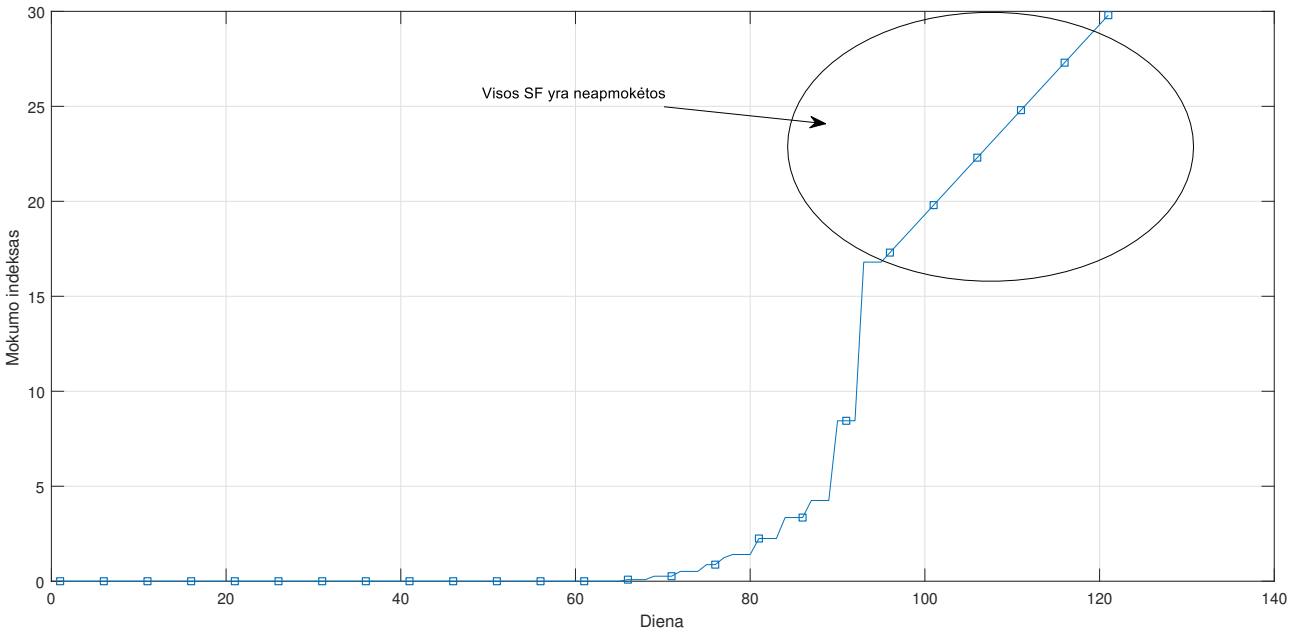
Parinkto sistemos išėjimo grafinė interpretacija pateikiama 3.1 paveiksle.



3.1 pav. Mokumo indekso grafinė interpretacija

Kol įmonės mokėjimų istorijoje nėra NSF, remiantis (3.1) formule, mokumo indeksas yra lygus nuliui. Atsiradus pirmai NSF šis santykis tampa teigiamas. Įmonei artėjant prie bankroto ribos nelieka ASF ir mokumo indekso reikšmė artėja prie begalybės. Parenkant modelį svarbu, jog šioje sistemoje prognozuotos reikšmės ir tikrosios reikšmės nesutapimas neturi prasmės. Čia svarbu, jog tikrosios mokumo indekso vertės kitimo tendencija būtų panaši į modelio kitimo tendenciją.

Remiantis (3.1) formule, galimas toks laiko tarpas, kaip įmonės istorijoje nelieka ASF, dėl to mokumo indekso vertė artėja prie begalybės. Modelyje ši situacija vaizduojama tiesiniu mokumo indekso kitimu. Grafinė aptariamo mokumo indekso kitimo interpretacija pateikiama paveiksle.



3.2 pav. Mokumo indekso kitimo laike grafinė interpretacija nelikus ASF

Pažymėtas tiesinis mokumo indekso kitimas parodo, jog įmonė nebeapmoka savo sąskaitų ir bankrutavo.

3.3. Modelio įėjimo duomenų formavimas

Sukurtoje mokumo indekso prognozavimo sistemoje modelio įėjimo duomenys suformuojami iš SF ir jų istorinių duomenų. Remiantis analizuota teorine medžiaga, buvo sudarytas 15-os požymių sąrašas, apibūdinantis SF ir įmonę bei jos istoriją [9]:

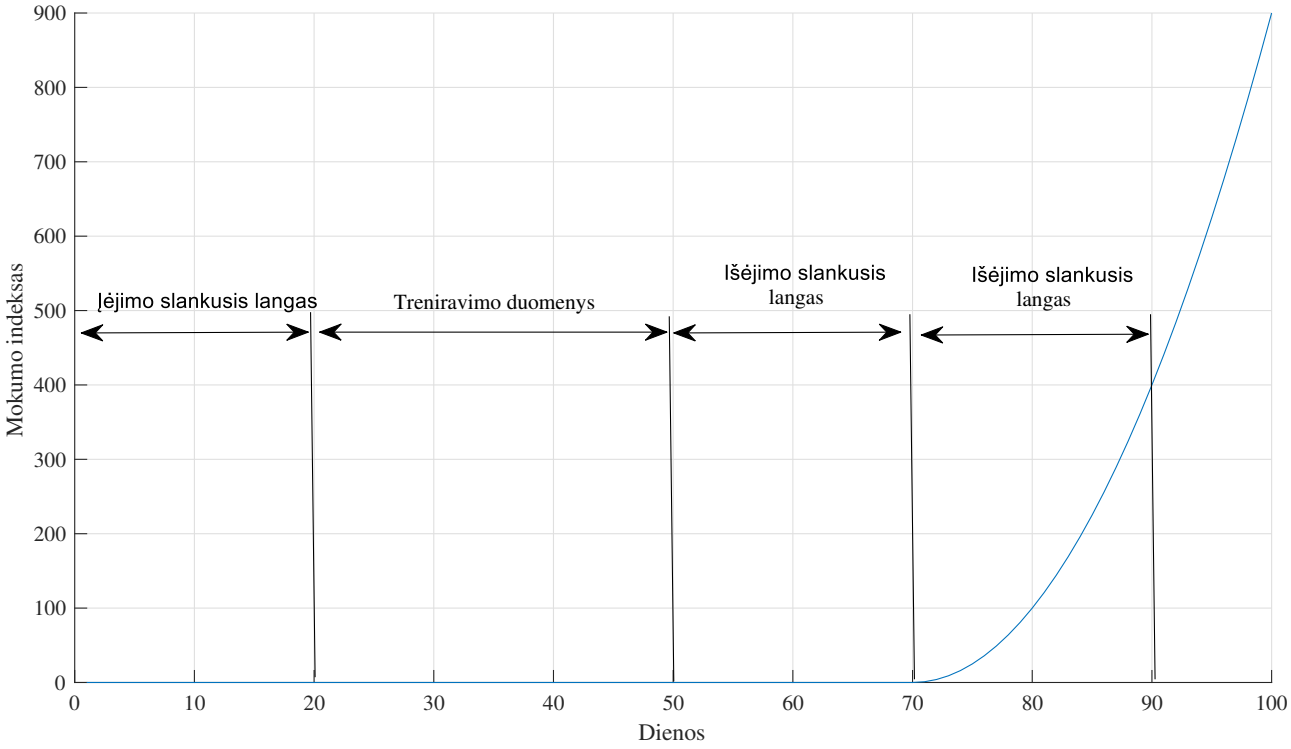
1. bendra SF suma;
2. maksimali SF suma;
3. minimali SF suma;
4. bendras SF kiekis;
5. SF kiekis per dieną;
6. vidutinis SF apmokėjimo vėlavimas;
7. standartinis SF apmokėjimo vėlavimo nuokrypis;
8. SF, apmokėtų pirmu mokėjimu, santykis su SF kiekiu;
9. neapmokėtų SF kiekio ir SF kiekio santykis;
10. SF, apmokėtų antru mokėjimu, santykis su SF kiekiu;

11. dažniausiai pasitaikantis mokėtojo tipas;
12. sektorius;
13. neapmokėtų SF kiekis;
14. apmokėtų SF kiekis;
15. apmokėtų ir visų SF kiekių santykis.

Sudarytas 15-os požymių sąrašas bus naudojamas ir klasifikavimo modelyje. Visi sąraše paminėti požymiai sudaromi iš istorinių SF duomenų konkrečiai įmonei. Laikotarpis, iš kurio apskaičiuojamos minėtos metrikos, yra vienas įėjimo duomenų slankusis langas. Regresijos modelyje šio lango trukmė buvo 30 kalendorinių dienų. Maksimali ar minimali SF suma, vyraujantis mokėtojo tipas ir sektorius priskiriami SF lygmeniui, bendras SF kiekis ar standartinis SF apmokėjimo vėlavimo nuokrypis gali būti priskirti antrajam požymių lygmeniui. Šioms metrikoms apskaičiuoti reikalingi SF istoriniai duomenys.

3.4. Modelio treniravimo duomenų paruošimas

Kompanijos mokumo problemų aptikimo modeliui apmokyti reikia paruošti treniravimo duomenis. Kadangi sudarant tiek modelio įėjimo, tiek modelio išėjimo duomenis naudojami slankieji langai, svarbu tinkamai parinkti treniravimo duomenis, jog būtų įmanoma apskaičiuoti tiek įėjimo, tiek išėjimo parametrus. Treniravimo duomenų paruošimo logika pateikiama 3.3 paveiksle.



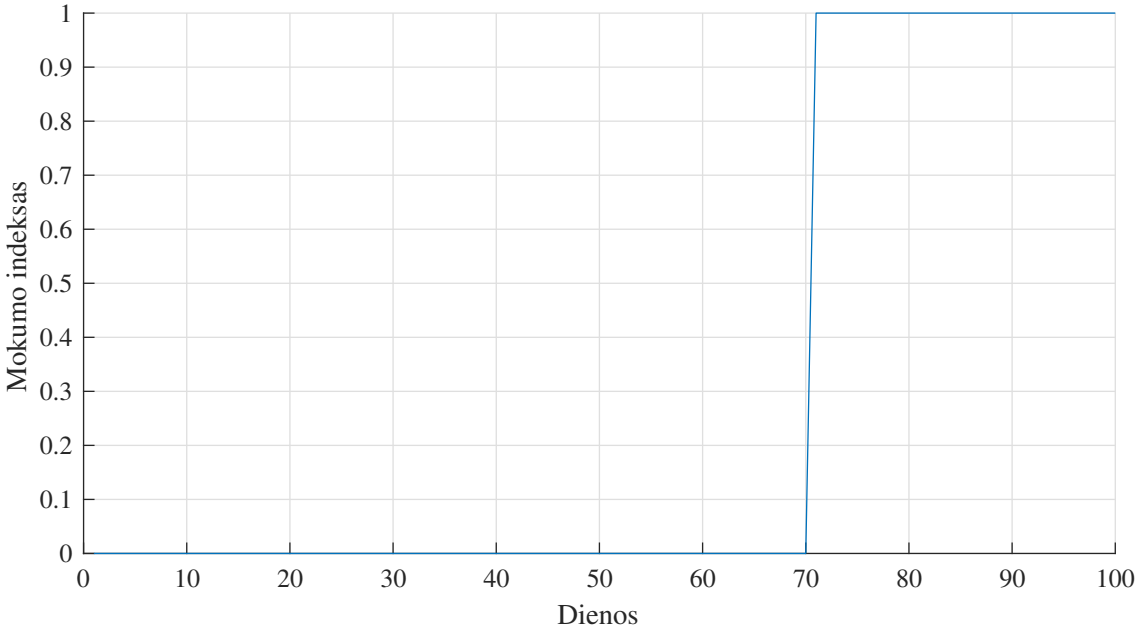
3.3 pav. Treniravimo duomenų paruošimas

Kaip matyti 3.3 paveiksle, treniravimo duomenų diapazonas priklauso nuo modelio įėjimo slankiojo lango, šiandieninės datos ir išėjimo slankiojo lango. Įėjimo ir išėjimo slankieji langai parodo, koks dienų laiko tarpas naudojamas sistemos įėjimo ir išėjimo vertėms sugeneruoti. Bakalauro baigiamajame projekte abu laikotarpiai buvo lygūs 30 dienų. Taigi sistemos treniravimo

duomenų imčiai sudaryti iš nagrinėjamos duomenų bazės išgaunami požymiai ir išėjimo vertė pradedant trisdešimtąja diena nuo istorinių duomenų pradžios ir baigiant 30 dienų iki dienos, kuriai norima atlikti mokumo indekso prognozę. Jei treniravimo duomenys prasidėtų anksčiau, nebūtų įmanoma apskaičiuoti visų požymių, reikalingų modelio įėjimo duomenims sudaryti. Jei treniravimo duomenų pabaigos data būtų vėlesnė, modelio apmokymui būtų naudojami duomenys, kurie bus panaudojami kaip įėjimo duomenys norimos dienos prognozei, taigi modeliui šie duomenys bus jau matyti. Su aptartais apribojimais suformuotas duomenų rinkinys naudojamas sistemai apmokyti.

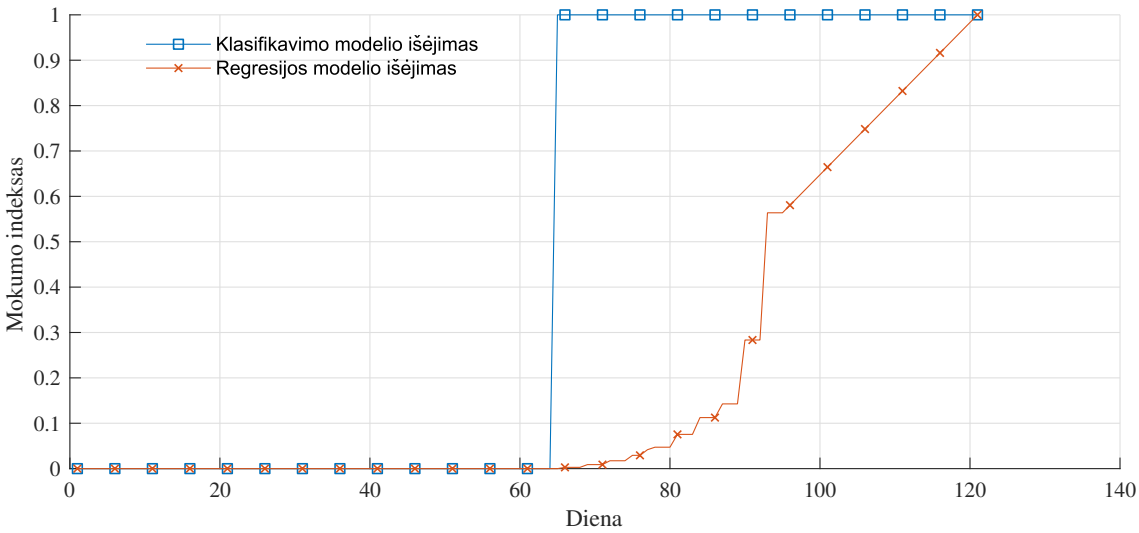
4. KLASIFIKAVIMO MODELIO KŪRIMAS

Remiantis praeitame skyriuje aptartu modelio išėjimo suformavimo principu sukuriamas klasifikavimo modelis, kurio išėjimas tampa binarinis, t. y. gali įgauti tik 0 arba 1 reikšmę. Kaip ir regresijos modelyje, sugeneruojamas mokumo indeksas, kuris šiuo atveju gali įgauti 0 reikšmę, jei per tiriamą laiko tarpą įmonės istorijoje neatsiras NSF, ir 1, jei per pasirinktą laiko tarpą įmonė neapmokės nors vienos SF. Grafinė šio kintamojo interpretacija pateikiama 4.1-ame paveiksle.



4.1 pav. Mokumo indekso grafinė interpretacija klasifikavimo atveju

Grafinis sukurtų klasifikavimo ir regresijos modelių mokumo indeksų palyginimas pateikiamas 4.2 paveiksle:

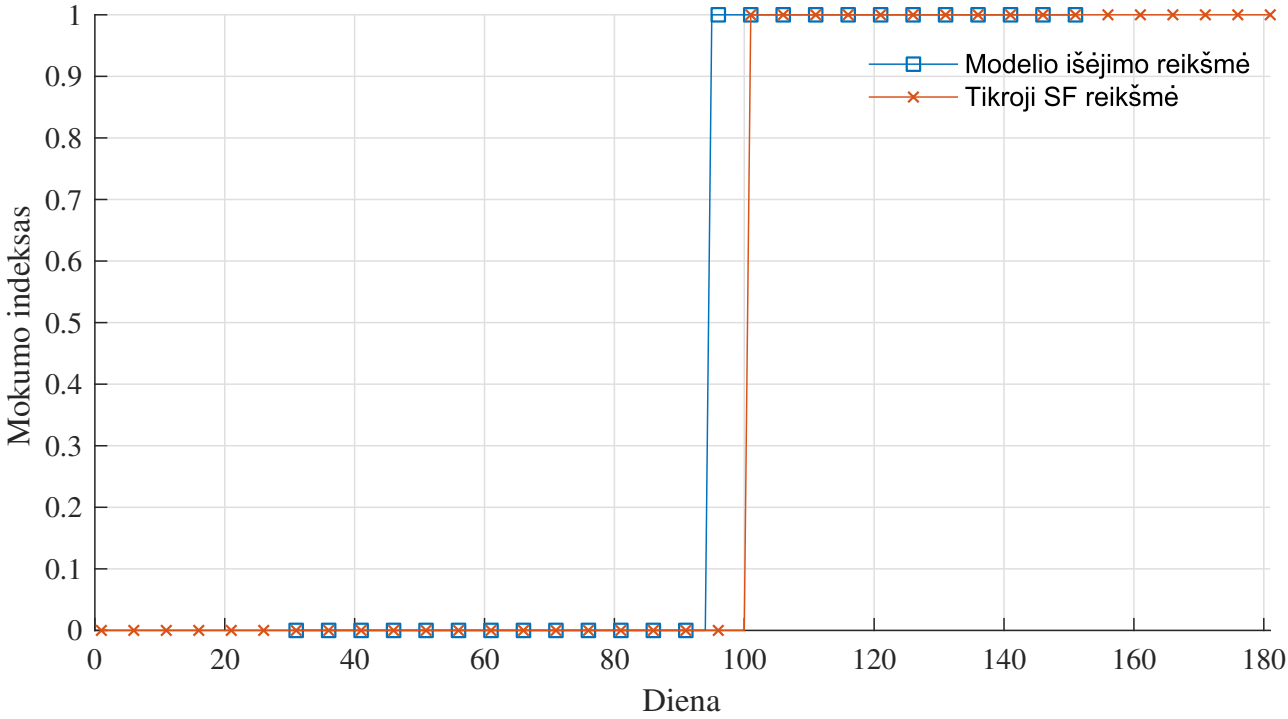


4.2 pav. Mokumo indeksas klasifikavimo ir regresijos atveju

Klasifikavimo modelis nepriklausomai nuo NSF kiekio grąžina tokią pačią išėjimo reikšmę, o regresijos modelio išėjimo reikšmė priklauso nuo NSF kiekio. Tiesinis šio parametro kitimas,

matomas nuo 97 dienos, parodo, jog įmonėje visos SF yra neapmokėtos, tuo tarpu klasifikavimo modelis tokios papildomos informacijos neteikia.

Nagrinėjamos įmonės istorijoje atsiradus NSF, mokumo indekso vertė pasiekia 1, taip parodydama, jog po pasirinktos išėjimo slankiojo lango trukmės įmonės istorijoje bus NSF. Jei modelio įėjimo ir išėjimo slankiųjų langų trukmės yra 30 dienų, sistemos prognozę galima atvaizduoti 4.3 paveikslu:



4.3 pav. Realus SF apmokėjimas ir kuriamo modelio prognozė

Kaip matyti paveiksle, modelio prognozuotų verčių yra mažiau. Tai atspindi 3.3 paveiksle pateiktą treniravimo duomenų parengimo logiką. Modelio išėjimo vertė pasiekia vienetą trisdešimčia dienų anksčiau, nei įmonės istorijoje pasirodo pirmoji tikroji NSF. Taigi banko atstovas jau prieš 30 dienų gali imtis veiksmų, nors realios NSF dar nėra.

4.1. Modelio kūrimas MATLAB aplinkoje

Suformavus modelio įėjimų ir išėjimų duomenis bei aptarus modelio treniravimo duomenų parinkimą kuriamas sistemos modelis MATLAB aplinkoje. Šiame projekte sukurti modeliai, paremti metodais:

1. klasifikavimo medžių kolektyvu;
2. Bajeso klasifikatoriumi;
3. atraminėmis vektorių mašinomis;
4. gilaus mokymosi neuroninio tinklu.

4.1.1. Klasifikavimo medžių kolektyvas

Klasifikavimo medžių kolektyvas remiasi prielaida, kad sujungus daug blogai besimokančių sprendimo medžių ir apmokius kaip vieną modelį, gaunamas geras prognozavimo rezultatas. MAT-

LAB aplinkoje šis metodas realizuojamas, panaudojant *fitcensenble* metodą, kuris aprašomas:

$$Mdl = fitcensemble(X, Y),$$

kur X – įėjimo duomenys, aprašantys dviejų klasių požymius, Y – reikšmės apibūdinančios, kuriai klasei priklauso įėjimų rinkinys (angl. *class labels*). Aptariamas metodas remiasi keleto klasifikavimo modelių sudarymo ir nagrinėjimo principu [1]. Nagrinėjama klasikinė mokymosi problema. Apmokamam modeliui sudaromi treniravimosi duomenys $\{(x_1, y_1), ..., (x_m, y_m)\}$ forma nežinomai funkcijai $y=f(x)$. x_i reikšmės tipiškai pateikiamos vektoriaus $\{x_{i,1}, x_{i,2}, x_{i,3}, ..., x_{i,n}\}$ pavidalu, kur kiekviena reikšmė apibūdina tam tikrą požymį. y reikšmės atitinka klases. Sudarius treniravimo imtį, mokymosi algoritmas apsimoko ir grąžina klasifikatorių. Modeliui davus naują x reikšmę, grąžinama atitinkama prognozuota y reikšmė [1]. Sudaromi keli klasifikavimo modeliai, kurie susiejami kokiu nors būdu, dažniausiai sprendimų svorio matrica. Tokiu būdu sprendimas priimamas atsižvelgiant į visus modelio narius ir iš kelių nepriklausomų rezultatų išrenkamas vienas. *fitcensemble* metodas pagal nutylėjimą naudoja logitbustinimo (angl. *LogitBoost*) metodą, kurio detalų veikimo paaiškinimą galima rasti MATLAB aplinkos techninėje dokumentacijoje [18]. Logitbustinimo algoritmas gali būti apibūdintas kaip pasirinkimo procesas, kuris pasirenka mažą rinkinį klasifikatorių su mažiausiomis paklaidomis ir jų svoriniais koeficientais [19]. Galutinis klasifikatorius yra laikomas stipriu, nes jis sudarytas iš kombinacijos silpnų klasifikatorių. Nors kiekvienas silpnas klasifikatorius negali suteikti gero klasifikavimo mokymo pavyzdžiams, tačiau tinkama svorinių koeficientų kombinacija su kitais klasifikatorių koeficientais gali pagerinti paskutinio klasifikavimo atlikimą.

4.1.2. Bajeso klasifikatorius

Bajeso metodas naudojamas klasifikuojant naujus duomenis, kai jų klasės nėra žinomos, todėl ieškoma labiausiai tikėtina klasė. Bajeso klasifikatorius remiasi tikimybių teorijoje naudojamomis Bajeso taisyklėmis. Turimi duomenys, kurių parametrai yra nepriklausomi bei vienodai lemia klasifikavimo rezultatą. Jie su klasėmis siejami C_l, l=1,...,k, kur k - klasių skaičius. Kiekvienas objektas X_i turi n nepriklausomų požymių reikšmių $x_{i1}, x_{i2}, ..., x_{in}$. Nustačius, kuriai klasei priklauso objektas, skaičiuojama aposteriorinė tikimybė:

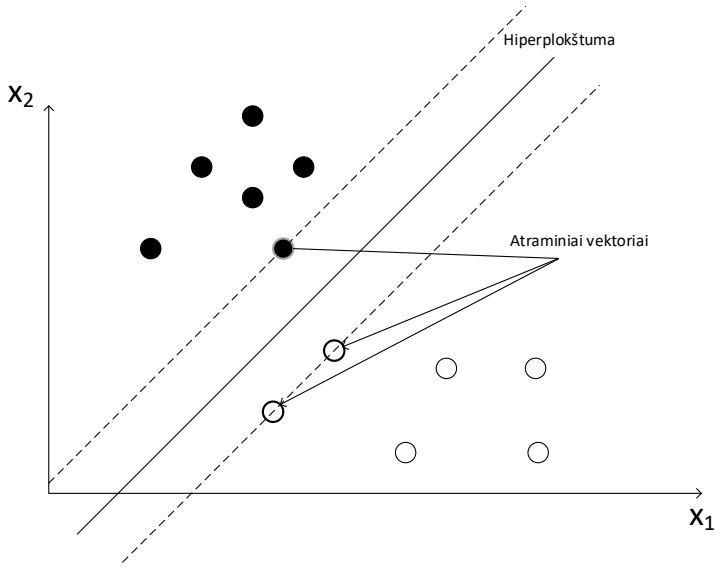
$$P(C_l|X_i) = \frac{P(X_i|C_i)P(C_i)}{P(X_i)}, \tag{4.1}$$

Šis dydis parodo, kokia tikimybė, jog X_i priklauso klasei C_i. Ši tikimybė apskaičiuojama visoms klasėms. Naujas duomenų objektas priskiriamas tai klasei, kurios aposteriorinė tikimybė didžiausia [20]. Aptartas metodas MATLAB aplinkoje realizuojamas *fitcnb* funkcija.

4.1.3. Atraminių vektorių mašina

Atraminių vektorių mašinos (angl. *Support Vector Mashine*) metodas remiasi struktūrinės rizikos mažinimo teorija [21]. Pagrindinė atraminių vektorių mašinos idėja yra įėjimo vektorių žemėlapio sudarymas, panaudojant kelių dimensijų savybių erdvę ir šioje erdvėje sukonstruojant

hiperplokštumą, kuri atskirtų klasių duomenis. Atraminių vektorių mašinos tikslas – maksimaliai sumažinti klasifikavimo paklaidą, maksimizuojant atstumą tarp skirtingų klasių duomenų ir sukurtos hiperplokštumos. Grafinė aptarto metodo interpretacija pateikiama 4.4 paveiksle.



4.4 pav. Atraminių vektorių mašinos grafinė interpretacija

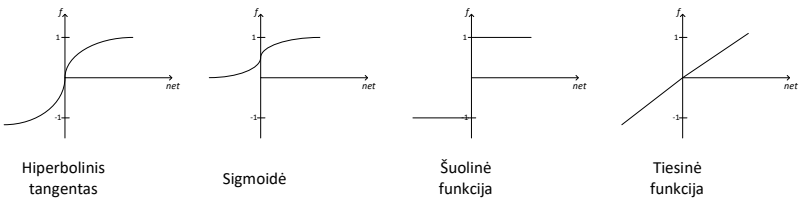
MATLAB aplinkoje šis metodas realizuojamas *fitcsvm* funkcija.

4.1.4. Gilaus mokymosi neuroninis tinklas

Dirbtinis neuroninis tinklas - tam tikrų netiesinių matematinių funkcijų rinkinys [15]. Šių tinklų struktūra apibūdinama kaip elementų, dar vadinamų neuronais, kurie sujungti tarpusavyje, visuma [22]. Kiekvieno neurono išėjimo signalas gali būti apskaičiuojamas panaudojant išraišką:

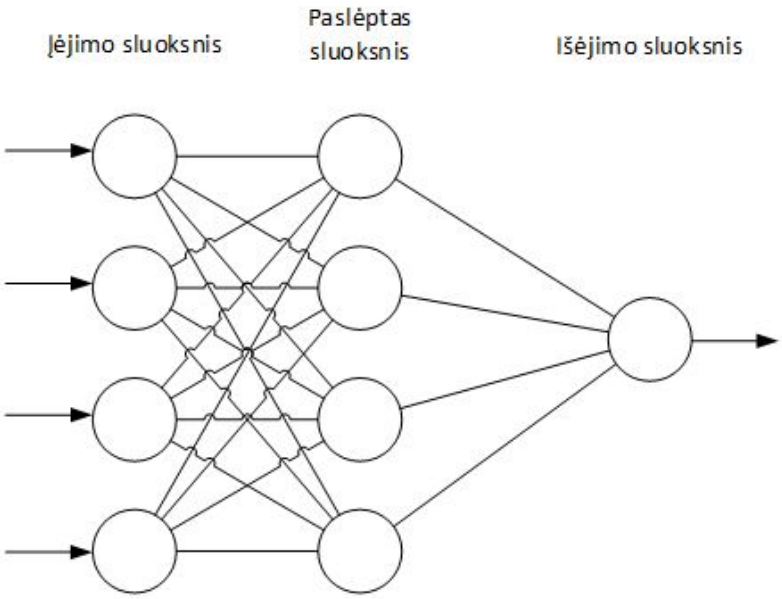
$$y = f(\sum_{i=1}^{n} w_i x_i + w_0) \tag{4.2}$$

kur $x_{1...n}$ neuronų įėjimo vertės, $w_{1...n}$ jungčių svoriai, w_0 - slenksčio reikšmė. Funkcija f vadinama neurono perdavimo funkcija. Dažniausiai pasitaikančios neurono perdavimo funkcijos pateikiamos 4.5 paveiksle.



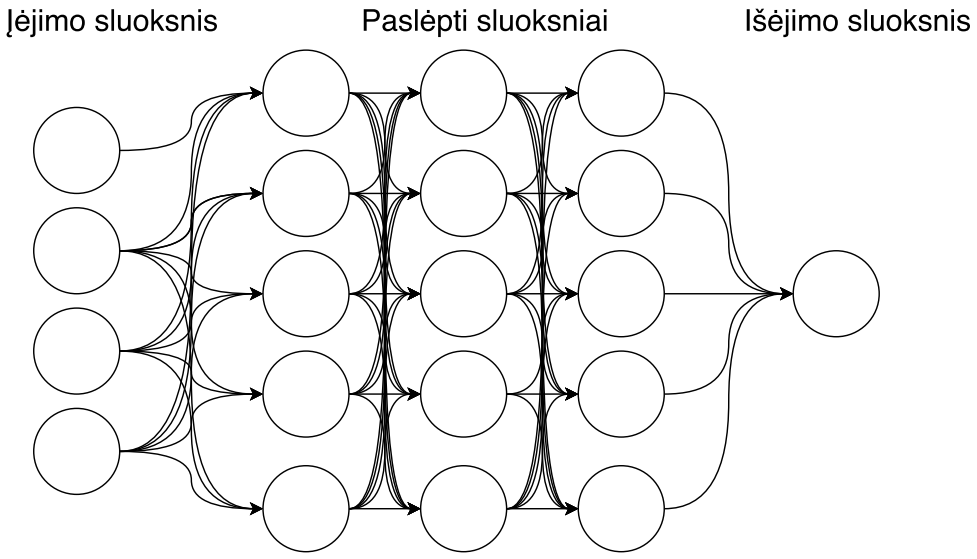
4.5 pav. Dažniausiai naudojamos neurono perdavimo funkcijos. Adaptuota iš [16]

Visuma minėtų neuronų turinčių tokią pačią perdavimo funkciją, laikoma sluoksniu. Būtent tokie sluoksniai jungiami tarpusavyje, taip sudarant dirbtinį neuroninį tinklą. Tokio neuroninio tinklo pavyzdys pateikiamas paveiksle.



4.6 pav. Neuroninio tinklo struktūros pavyzdys. Adaptuota iš [9]

Neuroniniai tinklai pradėti tobulinti dar 1943 metais, kai Warrenas McCullochas ir Walteris Pittsas sukūrė kompiuterinį modelį, paremtą slenksčių logika [23]. Proveržis neuroninių tinklų tobulinime įvyko 1975 metais pristačius atgalinės sklaidos algoritmą, kuris efektingai sprendė užduotis, keisdamas jungčių svorius, taip pagerindamas pirmtako rezultatus [24]. Vis dėlto, paminėti neuroniniai tinklai dažniausiai būdavo iš vieno paslėpto sluoksnio, kadangi tuometiniai kompiuteriai nebuvo pakankamai galingi apdoroti ir apmokyti didesnius neuroninius tinklus [25]. Jei neuroninis tinklas susideda iš dviejų arba daugiau paslėptų sluoksnių, jis yra laikomas gilaus mokymosi neuroniniu tinklu [25]. Galima tokio tinklo struktūra pateikiama 4.7 paveiksle:



4.7 pav. Gilaus mokymosi neuroninio tinklo struktūros pavyzdys. Adaptuota iš [25]

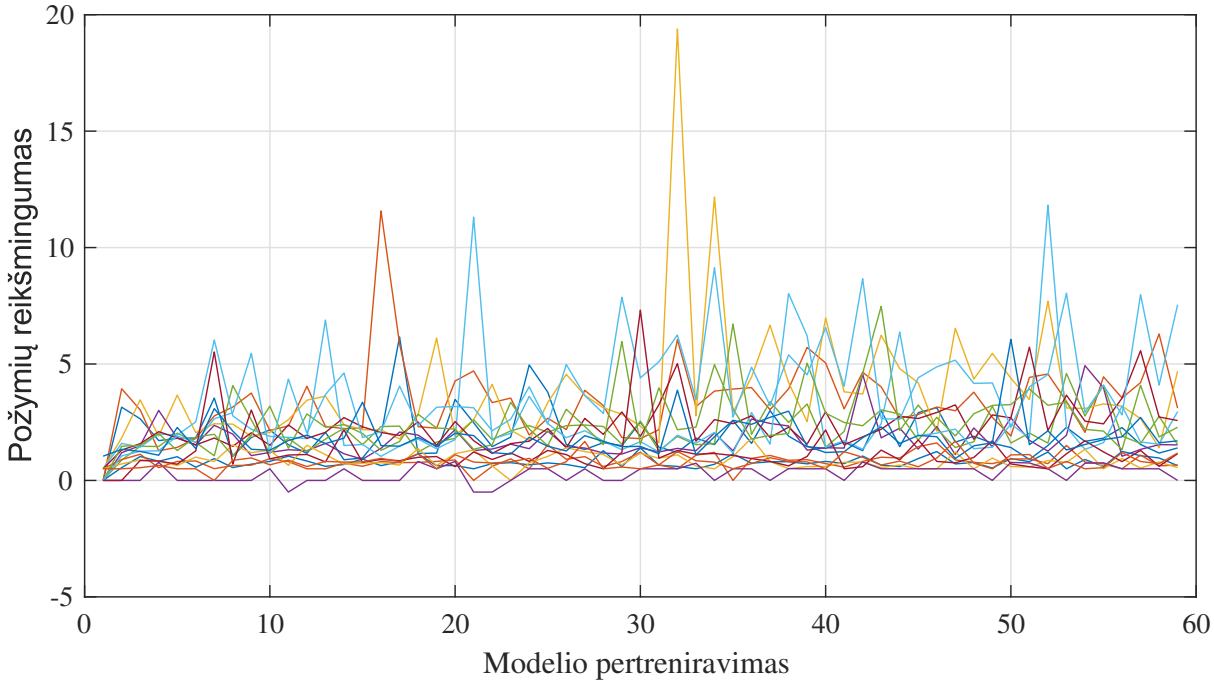
Šį tinklą sudaro įėjimo sluoksnis, trys paslėpti sluoksniai ir vienas išėjimo sluoksnis. Nuo 2011 metų išpopuliarėjus konvoliuciniams neuroniniams tinklams bei patobulėjus kompiuterinėms technologijoms gilaus mokymosi neuroniniai tinklai pradėti plačiai taikyti vaizdo ar kalbos atpažinime [26, 27, 28]. Nagrinėjant didelius kiekius duomenų gilaus mokymosi neuroniniai tinklai tikslumu konkuruoja su klasikiniais mašininio mokymo modeliais [29]. Nagrinėjant įmonės kredito riziką gilaus mokymosi neuroniniai tinklai nesugebėjo nukonkuruoti tradicinio bustinimo (angl. *boosting*) metodo [25].

5. REIKŠMINGIAUSIŲ POŽYMIŲ NUSTATYMAS

Tiek sukurtame regresijos modelyje, tiek nagrinėtoje literatūroje, modeliui apmokyti buvo naudojamas požymių sąrašas, kurį sudaro SF ir istorinė kliento informacija [5, 6, 7, 8]. Šių požymių apskaičiavimas reikalauja labai daug laiko, apmokant modelį treniravimosi stadijoje, todėl siekiama optimizuoti šių požymių panaudojimą ir jo apmokymui naudoti tik reikšmingiausius požymius.

5.1. Požymių reikšmingumo identifikavimas

Panaudojant požymių svarbumo nustatymo funkciją (angl. *PredictorImportance*) identifikuojamas kiekvieno požymio reikšmingumo kitimas modelio pertreniravimo metu. Modeliui apmokyti naudojamų 15 požymių reikšmingumo kitimas pateikiamas 5.1 paveiksle:

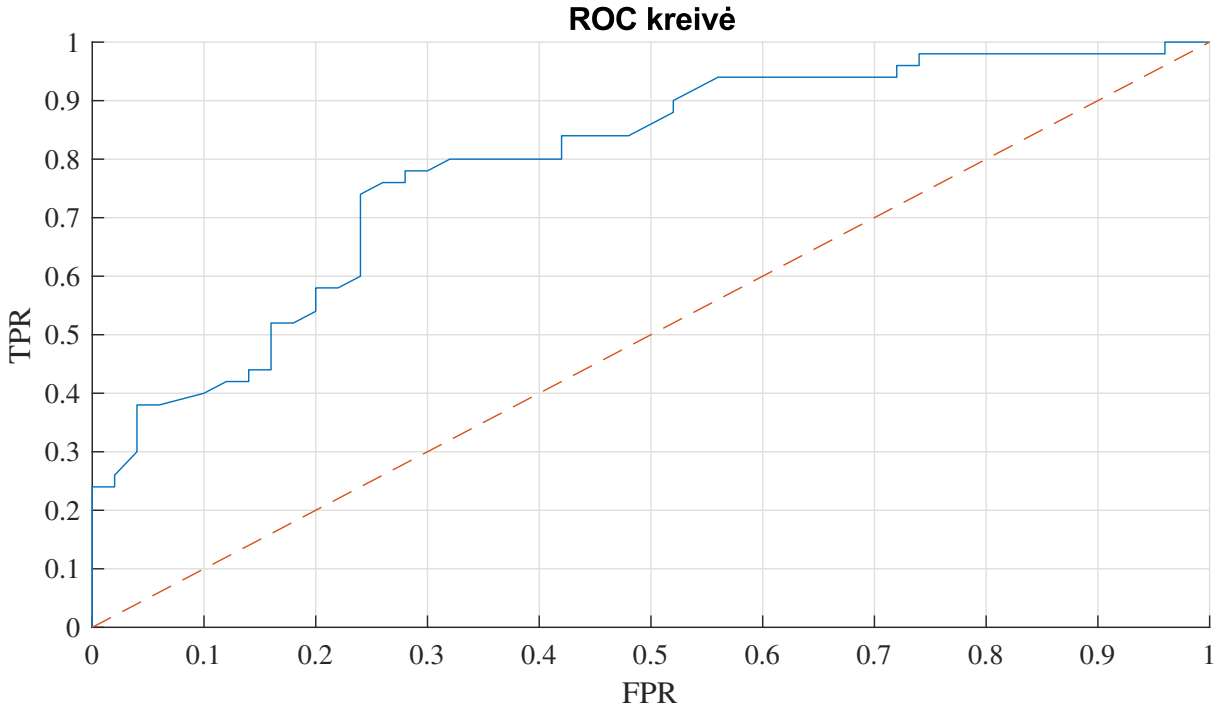


5.1 pav. Požymių reikšmingumo kitimas

Grafike matyti, jog kiekvieno modelio apmokymo metu požymių reikšmingumas yra skirtingas ir kinta ribose nuo 0 iki 20. Iš gautų požymių reikšmingumų negalima atrinkti vieno ar kelių reikšmingiausių požymių, taigi toliau atliekamas tyrimas, siekiant nustatyti optimalų slenkstį, kurio pagalba būtų atrenkami ir į mokymo imtį įtraukiami tik svarbiausi požymiai. Šiam slenksčiui nustatyti pasitelkiamos ROC (angl. *Receive roperating characteristic*) kreivės.

5.2. ROC kreivė

ROC kreivė – grafikas, rodantis klasifikatoriaus jautrumo ir specifiškumo (tiksliau, vieneto ir specifiškumo skirtumo) sąryšį [30]. ROC kreivės pavyzdys pateikiamas 5.2 paveiksle.

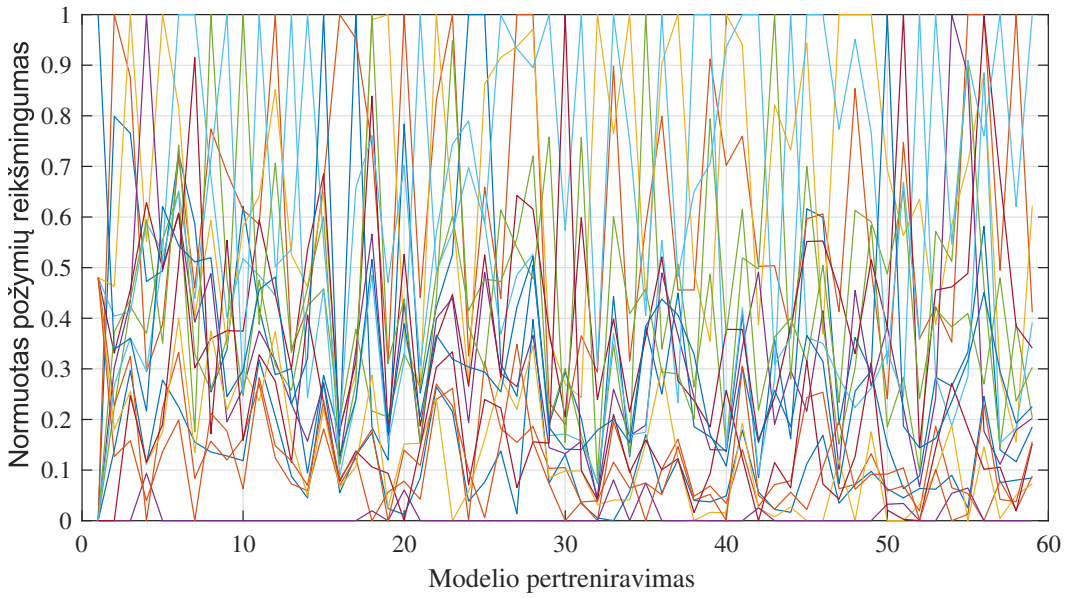


5.2 pav. ROC kreivės pavyzdys

Kaip matyti 5.2 paveiksle, ROC kreivės abscisių ašyje atvaizduojamas FPR (angl. *False positive rate*) – skirtumas tarp 1 ir modelio specifiškumo, o ant ordinačių ašies – TPR (angl. *True positve rate*), atspindintis modelio jautrumą. Kiekvienai galimai šio slenksčio reikšmei, abiejose ašyse atvaizduojama teigiamų klasifikatoriaus spėjimų dalis (vertikalioje – teisingų teigiamų, horizontalioje – klaidingų teigiamų). Esant didelėms slenksčio reikšmėms, klasifikatorius beveik ar ir visiškai nepateikia teigiamų spėjimų. Mažinant slenkstį, abiejų spėjimų dalis priartėja prie vieneto. Visiškai atsitiktinai išvadą spėjančio analizatoriaus ROC kreivė yra įstrižai grafiką kertanti tiesė (parodyta punktyrine linija). Analizatorius tuo geresnis, kuo aukščiau šios tiesės yra jo kreivė. Šią ROC kreivės savybę taip pat galima išreikšti AUC (angl. *Area under curve*), ploto po kreive, parametru. Kuo AUC arčiau vieneto, tuo klasifikatorius laikomas geresniu [30].

5.3. Optimalaus slenksčio nustatymas

ROC kreivių ir ploto po jomis metodika panaudojama optimaliam požymių reikšmingumo slenksčiui nustatyti. 5.1 -ajame paveiksle matyti, jog požymių reikšmingumas kinta ribose nuo 0 iki 20, tačiau daugumos požymių reikšmingumas kinta tik nuo 0 iki 5. Ieškoti optimalaus slenksčio visame reikšmingumo intervale nėra prasmės, nes daugelyje atvejų, ribose nuo 5 iki 20, į treniravimo duomenų imtį nepateks nei vienas požymis. Todėl kiekvienoje X ašies atskaitoje požymių reikšmingumai normalizuojami 0 ir 1 ribose. Tokiu atveju, keičiant slenkstį, visada bent vienas požymis bus įtrauktas į mokymo imtį. Normalizuotų požymių reikšmingumo kitimas pateikiamas 5.3 paveiksle.

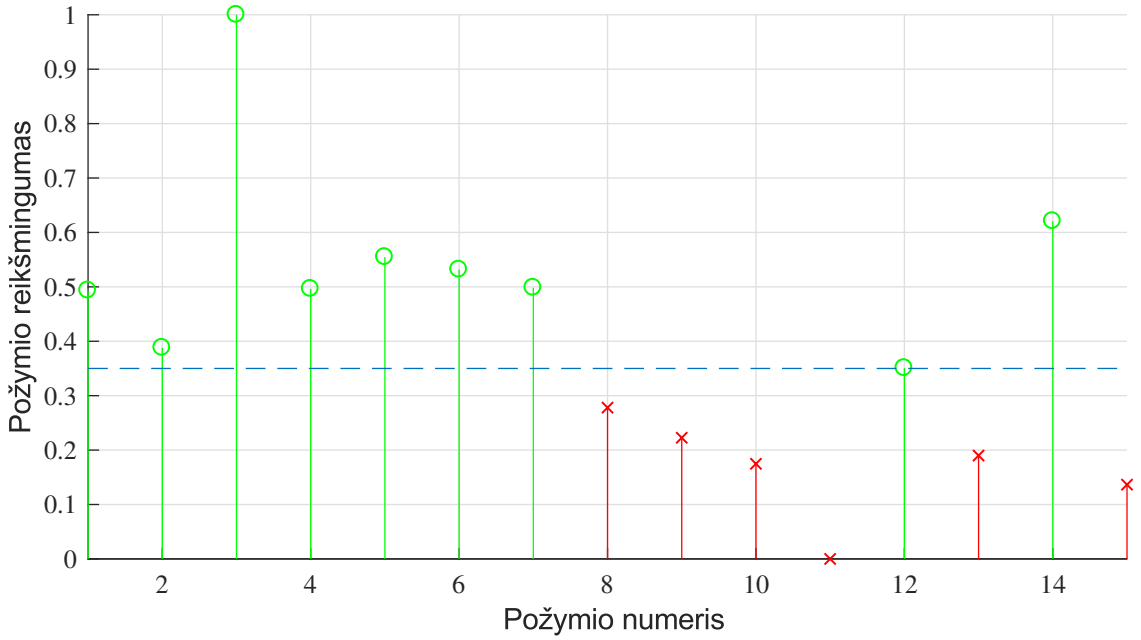


5.3 pav. Normuotų požymių reikšmingumo kitimas

Naudojant normalizuotą požymių reikšmingumą ir keičiant slenksčio ribą nuo 0 iki 1, kiekvienos iteracijos metu į mokymo imtį bus įtrauktas bent vienas požymis. Požymių reikšmingumo slenkstis keičiamas 20 kartų, kiekvienoje iteracijoje slenksčio vertę padidinant 0,05. Apskaičiuojamos analizuojamo atvejo TPR ir FPR vertės, kurios naudojamos ROC kreivės sudarymui. MATLAB aplinkoje tai įgyvendinama panaudojant ROC kreivės sudarymo funkciją (angl. *perfcurve*):

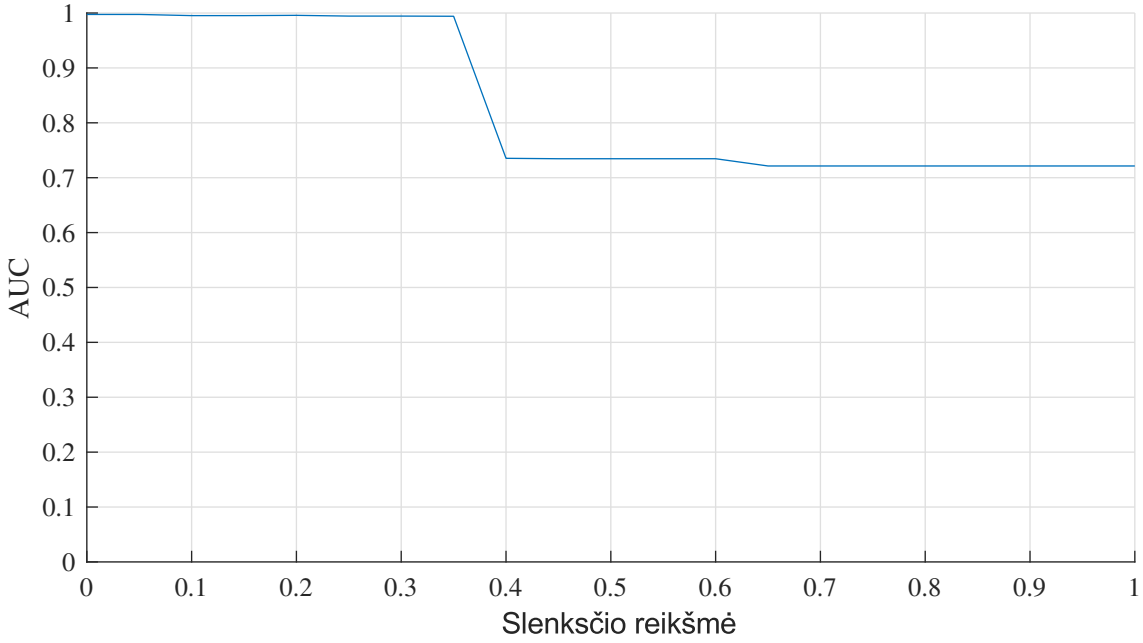
[X, Y, T, AUC] = perfcurve(labels, scores, posclass)

Svarbu paminėti, jog *perfcurve* metodas apskaičiuoja ROC parametrus panaudojant tik modelio treniravimosi metu naudojamus visų įmonių duomenis. Sudarant treniravimo įėjimo duomenis į imtį įtraukiami tik tie duomenys, kurie viršija nustatytą slenksčio reikšmę:



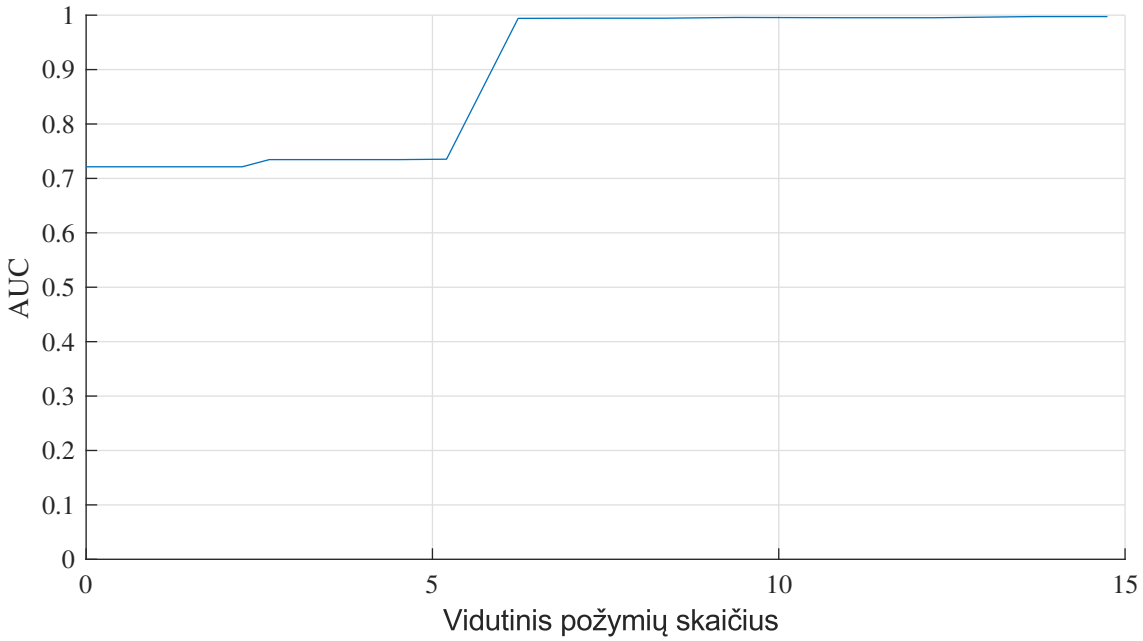
5.4 pav. Reikšmingų požymių atrinkimas panaudojant slenkstį

Kaip matome grafike, į mokymų imtį šiuo atveju įtraukiami tik požymiai, kurių reikšmingumas yra didesnis nei užduotas 0,35 slenkstis, šie požymiai grafike pažymėti rutuliuko simboliu. Kiekvienai gautai ROC kreivei apskaičiuojamas AUC parametras. Jo priklausomybė nuo slenksčio vertės pateikiama 5.5 paveiksle:



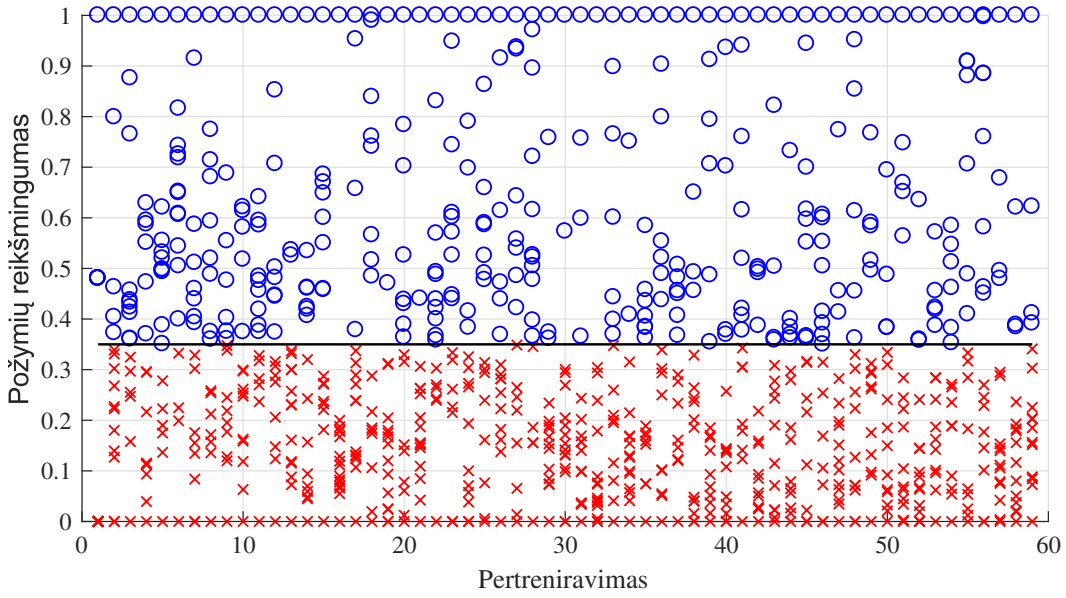
5.5 pav. Ploto po ROC kreive priklausomybė nuo slenksčio

Iš 5.5 duomenų matyti, jog perkopus 0,35 slenksčio vertę klasifikatorius žymiai pablogėja, tačiau bet kokiu atveju AUC vertė yra didesnė nei 0,5, taigi klasifikatoriaus tikslumas geresnis nei atsitiktinio spėjimo. AUC kitimas, priklausantis nuo naudojamų požymių skaičiaus, pateikiamas 5.6 paveiksle.



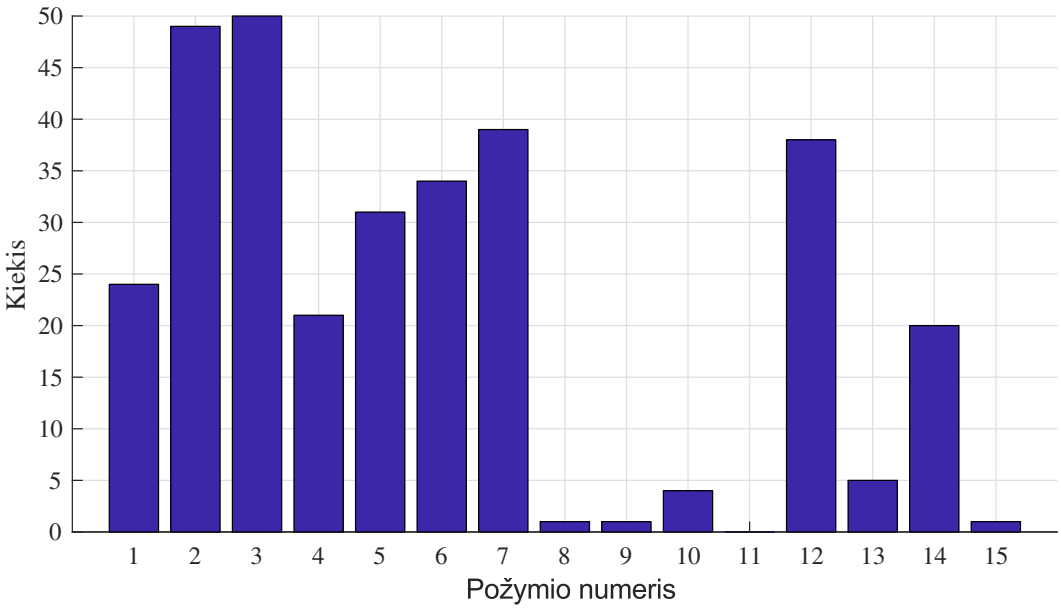
5.6 pav. Ploto po ROC kreive priklausomybė nuo požymių skaičiaus

Grafike matyti, jog naudojant daugiau nei 6 požymius AUC vertė artėja link 1. Remiantis šiais rezultatais, optimaliu slenksčiu pasirenkamas 0,35. Taigi parenkant treniravimo duomenis, modelio persimokymui atrenkami tik tie požymiai, kurių reikšmingumas didesnis nei 0,35. Tokiu būdu modelis neapkraunamas nereikšmingais požymiais, taip sutrumpinant modelio apmokymo trukmę. Siekiant identifikuoti 6 požymius, kurie modeliui apmokyti buvo reikšmingiausi, atrenkamos atskaitos, kurios kiekvieno persitreniravimo metu viršijo 0,35 vertės ribą. Grafinė šios atrankos interpretacija pateikiama 5.7 paveiksle.



5.7 pav. Reikšmingiausių požymių nustatymas panaudojant 0,35 slenkstį. Požymių reikšmingumas normuotas ribose nuo 0 iki 1

Į tolimesnius skaičiavimus įtraukiami požymiai, kurie viršijo juoda linija pažymėtą ribą. Suskaičiuojama, kiek kartų kiekvienas požymis buvo įtrauktas į šią imtį per 60 modelio persitreniravimų. Rezultatai pateikiami 5.8 histogramoje:



5.8 pav. Reikšmingiausių požymių histograma

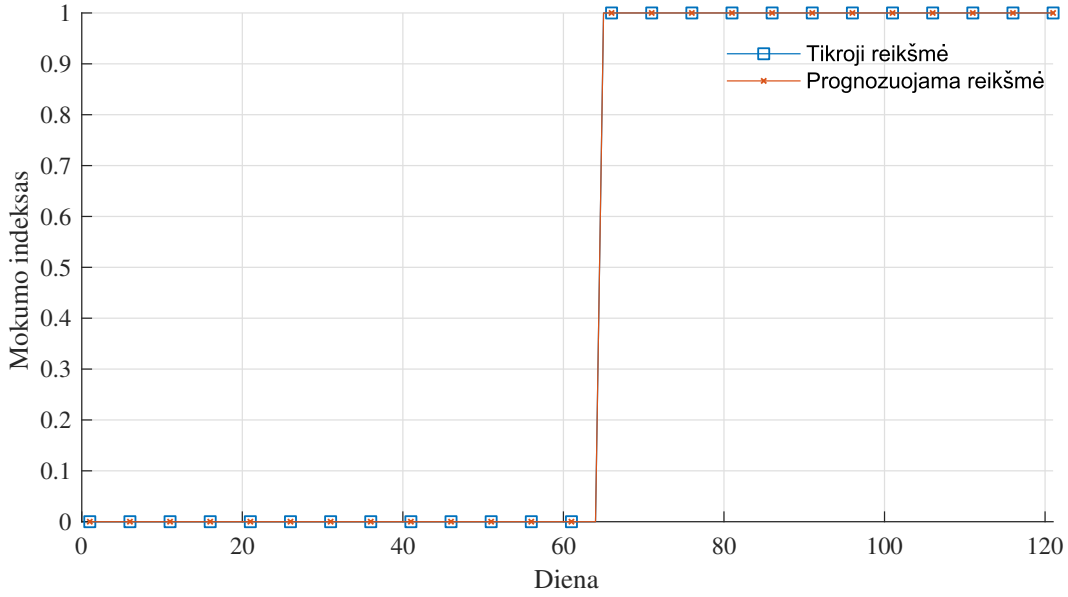
Iš histogramos duomenų matyti, jog daugiausiai kartų reikšmingiausi buvo šie požymiai:

1. maksimali SF suma;
2. minimali SF suma;
3. SF kiekis per dieną;
4. vidutinis SF apmokėjimo vėlavimas;
5. standartinis SF apmokėjimo vėlavimo nuokrypis;
6. sektorius.

Paminėti 6 požymiai laikomi reikšmingiausiais ir bus naudojami apmokant gilaus mokymosi neuroninį tinklą, aprašytą 8 skyriuje.

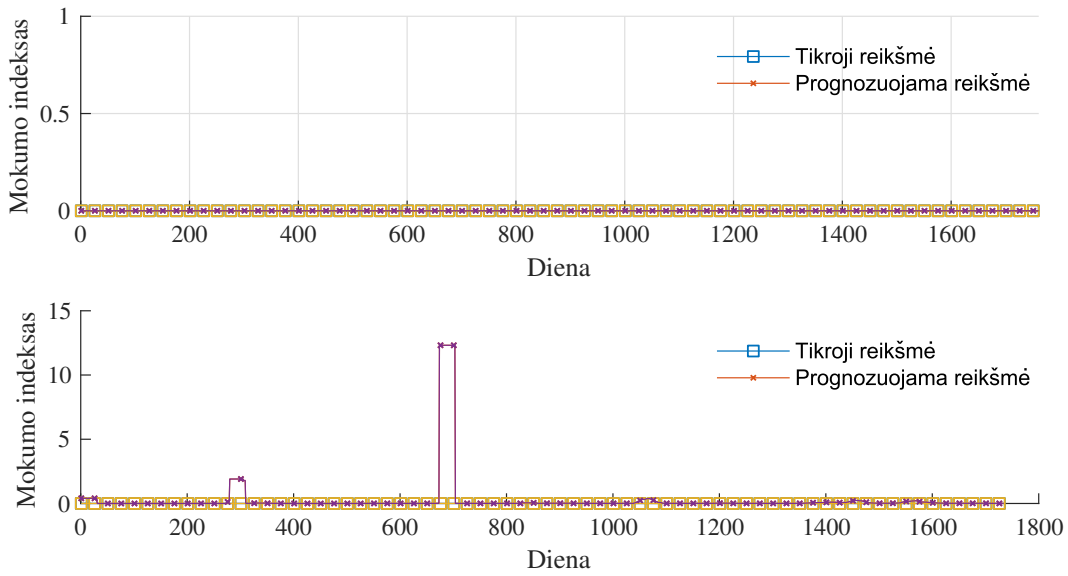
6. PROGNOZAVIMO REZULTATŲ APŽVALGA

Sukūrus ir apmokius klasifikavimo medžių kolektyvo modelį, atliekama įmonės mokumo prognozė. Atsitiktiniu būdu atrenkama viena įmonė, kurios istorijoje yra NSF. Mažmeninės prekybos pirmosios įmonės mokumo indekso prognozė pateikiama 6.1 paveiksle.



6.1 pav. Įmonės mokumo prognozė naudojant sprendimo medžių kolektyvo modelį

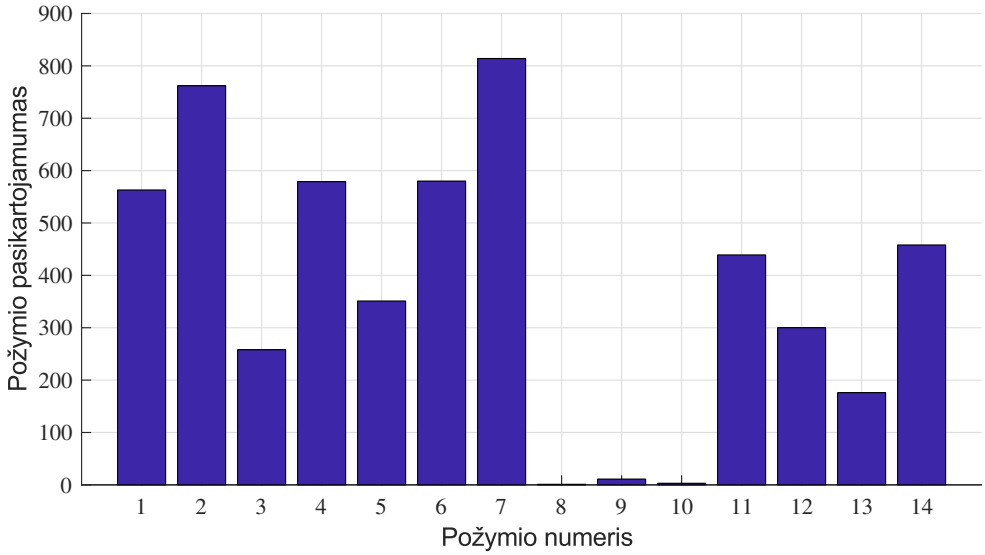
Iš gautų rezultatų matyti, jog modelio prognozuojama reikšmė visiškai sutampa su tikrąja mokumo indekso reikšme. Sukurtas modelis, kaip ir tobulinamas regresijos modelis, aptinka NSF prieš 30 dienų iki jos atsiradimo. Toliau nagrinėjama atsitiktinai atrinktos įmonės, kurios istorijoje nėra NSF, mokumo indekso prognozė. Klasifikavimo ir regresijos modelių prognozavimo rezultatų lyginimas pateikiamas 6.2 paveiksle:



6.2 pav. Įmonės mokumo prognozė klasifikavimo (viršuje) ir regresijos (apačioje) atveju

Nagrinėjant klasifikavimo atvejį, modelio ir tikrosios mokumo indekso vertės visiškai sutapo, o regresijos modelis klaidingai prognozavo NSF atsiradimą. Šiam reiškiniui ištirti atliekamas

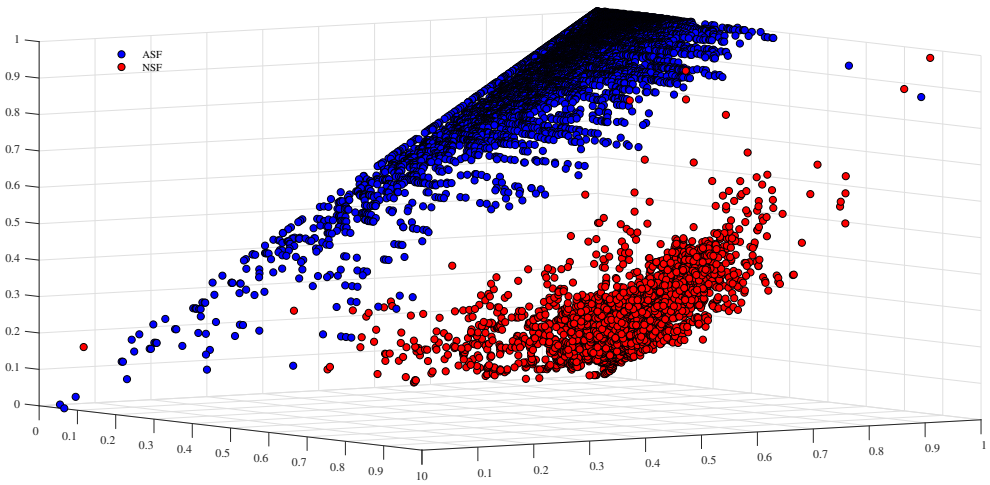
papildomas tyrimas, siekiant nustatyti, ar šiuo laiko momentu įmonės mokumo prognozavimui naudojami požymiai buvo artimi nemokios įmonės požymiams. Tada apskaičiuojami trys reikšmingiausi požymiai, siekiant atvaizduoti jų pasiskirstymą erdvėje. Pirmiausia suskaičiuojama, kiek kartų požymis per visą treniravimosi ciklą yra įtraukiamas į reikšmingiausių požymių trejetuką. Trys požymiai su didžiausiais pasikartojimo kiekiais atrenkami atvaizduoti trimatėje erdvėje. Grafinė šios paieškos interpretacija pateikiama 6.3 paveiksle:



6.3 pav. Svarbiausių požymių histograma

Iš histogramos matyti, jog septintas ir antras požymiai dažniau pasitaikė tarp reikšmingiausių požymių. Ketvirtas ir šeštas požymiai į reikšmingiausių požymių imtį buvo įtraukti atitinkamai 579 ir 580 kartų, taigi maksimali SF suma (2), vidutinis SF apmokėjimo vėlavimas (6) ir standartinis SF apmokėjimo vėlavimo nuokrypis (7) naudojami kaip reikšmingiausi požymiai šiame tyrime. Kadangi šių reikšmių kitimo diapazonai labai skirtingi, reikšmės normalizuojamos [0,1] ribose.

Norint nustatyti, ar mokumo indekso staigus išaugimas signalizuoja apie įmonės požymių panašumą į nemokios įmonės požymius, šios dvi klasės pavaizduojamos 3D erdvėje, kurios ašyse atidedamos požymių vertės. Klasių pasiskirstymas 3D erdvėje pateikiamas 6.4 paveiksle.

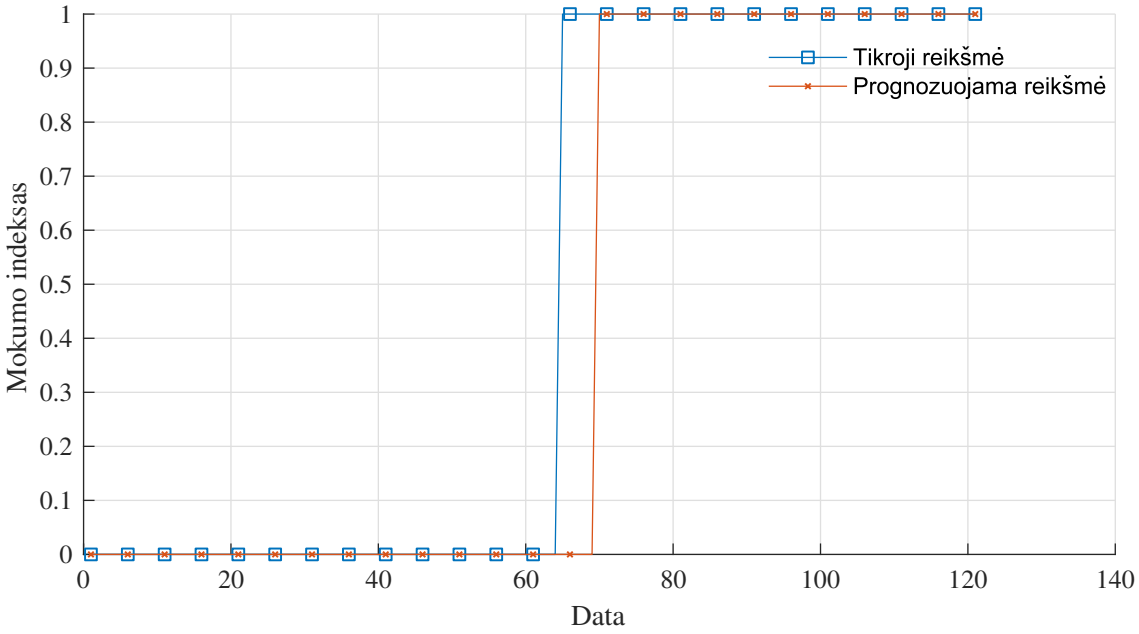


6.4 pav. ASF ir NSF pasiskirstymas erdvėje pagal požymių reikšmes

Kaip matyti, abi klasės erdvėje atsiskyrusios, taigi galima laikytis prielaidos, jei SF požymių taškas staigiai išaugus mokumo indeksui yra arčiau NSF taškų centro, tai tuo metu įmonės požymiai buvo artimi nemokios įmonės, tačiau įmonės finansiniai rodikliai stabilizavosi ir ji sugebėjo laiku apmokėti SF. Nagrinėtas atvejis pasitaikė 80 kartų per visą duomenų imtį ir daugiau nei 76% atvejų įmonės požymiai buvo panašūs į nemokių įmonių požymius, kadangi taškai buvo arčiau NSF klasės centro. Likusiais atvejais atstumų skirtumas buvo artimas 0, todėl negalima įvertinti, ar šiais atvejais modelis vienareikšmiškai klydo.

6.1. Atraminių vektorių mašinos modelio prognozavimo rezultatai

Toliau pateikiama mokumo indekso prognozė panaudojant atraminių vektorių mašinos modelį. Šio modelio prognozė mažmeninės prekybos pirmajai įmonei, kaip ir 6.1 pavaizduotame paveiksle, pateikiama 6.5 paveiksle:

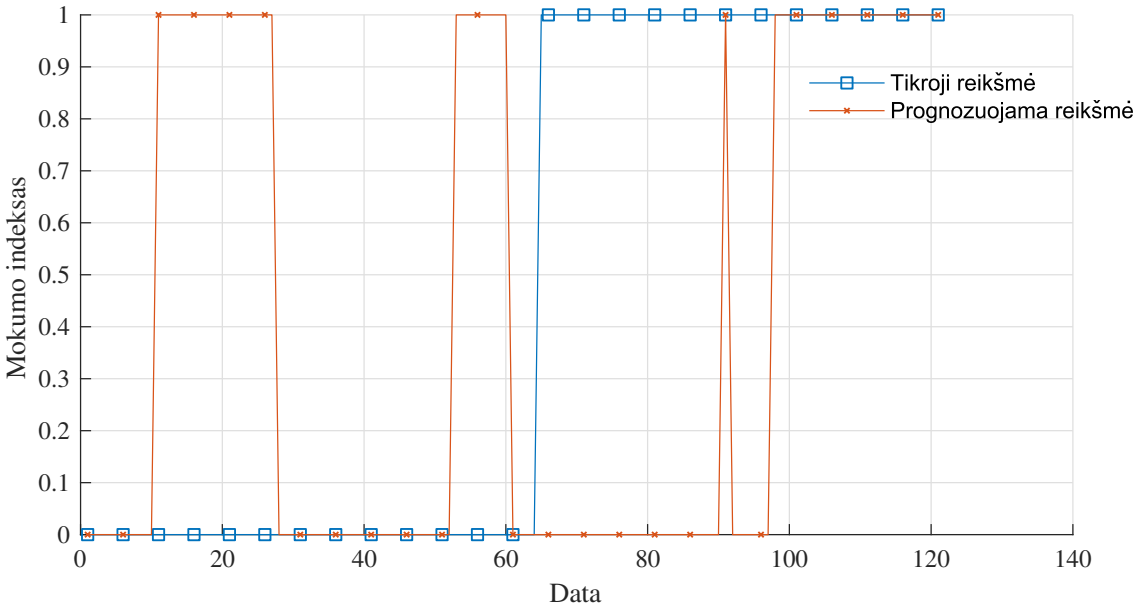


6.5 pav. Įmonės mokumo prognozė naudojant atraminių vektorių mašinos modelį

Kaip matyti grafike, šis modelis taip pat didžiąją dalį taškų prognozavo teisingai, tačiau nesutapo pirma reikšmė, kada mokumo indeksas pasiekė 1, taigi atraminių vektorių mašina paremtas modelis viena diena vėliau aptiko neapmokėtas SF. Tiriant įmonę, kurios istorijoje nėra NSF, atraminių vektorių mašimos modelis, kaip ir sprendimų medžių kolektyvo modelis, neklydo.

6.2. Bajeso modelio prognozavimo rezultatai

Taip pat atlikta prognozė panaudojant Bajeso klasifikatorių. Šio metodo prognozavimo rezultatai mažmeninės prekybos pirmajai įmonei pateikiami 6.6 paveiksle.



6.6 pav. Įmonės mokumo prognozė naudojant Bajeso modelį

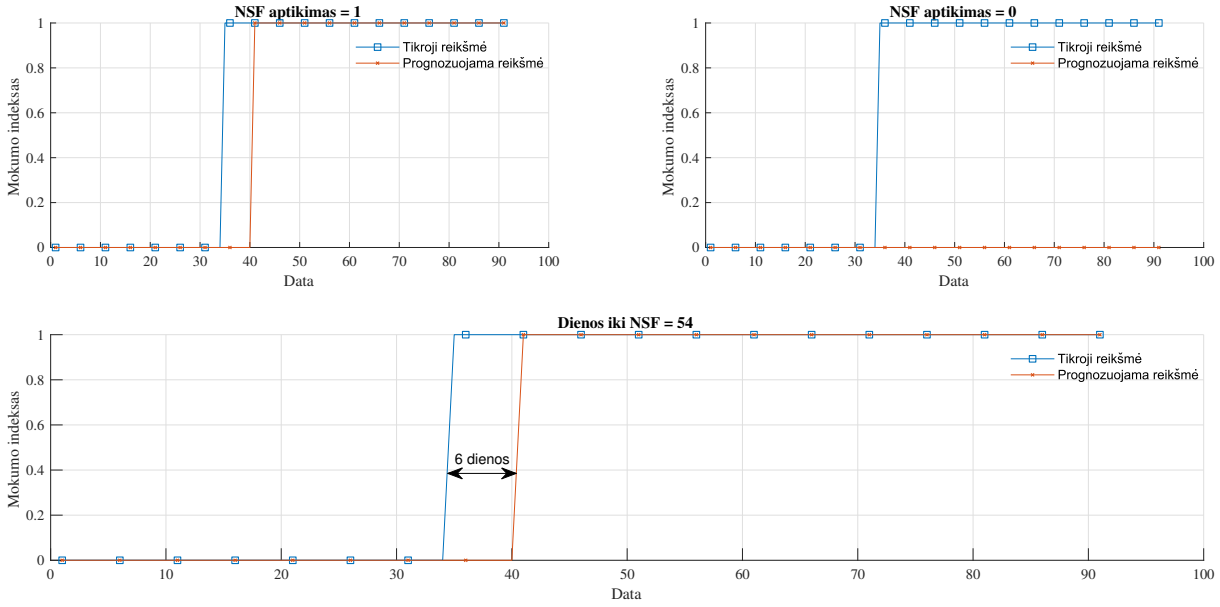
Kaip matyti grafike, prognozuota reikšmė daugeliu atveju nesutapo su tikrąja mokumo reikšme. Modelis prognozavo NSF įmonės istorijos pradžioje, kai dar NSF nebuvo, o atsiradus neapmokėtoms SF jas aptiko daug vėliau, nei prieš tai aptarti modeliai. Atlikus mokumo prognozę įmonėms, kurių istorijoje nėra neapmokėtų sąskaitų faktūrų, Bajeso modelis klaidingų alermų nepateikė, taigi prognozuotos ir tikrosios mokumo indekso vertės sutapo per visą prognozavimo laikotarpį.

7. NSF DETEKTAVIMO KOKYBĖS VERTINIMAS

Tiek regresijos modelyje (įgyvendintame bakalauro baigiamojo projekto metu), tiek sukurtame klasifikavimo modelyje, prognozuojama ar per nustatytą laiko tarpą įmonėje atsiras NSF. Analizuojamu atveju šis nustatytas laiko tarpas siekė 30 kalendorinių dienų. Toks pat laikotarpis taikomas ir apskaičiuojant modelio įėjimo kintamuosius. Siekiant kuo anksčiau nustatyti NSF, galima šiuos parametrus keisti. Įvertinant sukurtos sistemos prognozavimo kokybę buvo pasirinkti du ją apibūdinantys kriterijai:

1. įmonės NSF aptikimas;
2. dienų skaičius iki pirmos tikros NSF, modeliui signalizavus apie jų atsiradimą.

Pirmasis parametras yra binarinis ir parodo, ar modelis identifikavo NSF įmonės istorijoje nepriklausomai nuo laiko, kada buvo signalizuota apie NSF. Antrasis parametras parodo, prieš kiek dienų iki NSF atsiradimo modelis signalizuoja apie mokumo indekso padidėjimą. Kuriamame klasifikavimo modelyje ypač svarbu, jog sistema signalizuotų apie NSF atsiradimą ateityje. Klaidingas aliarmas nagrinėjamu atveju reikštų tik papildomą darbą banko analitikui, kuris turėtų tikrinti, ar įmonė moki, tačiau laiku nepastebėtos NSF gali sukelti daug didesnių finansinių nuostolių. Grafinis šių kriterijų paaiškinimas pateikiamas 7.1 paveiksle.



7.1 pav. Modelio kokybės kriterijų grafinis pavyzdys

Pirmajame grafike atvaizduojamos tikrosios ir modelio prognozuotos reikšmės. Pirmasis kriterijus analizuojamu atveju lygus 1, nes bent viena modelio prognozuota reikšmė lygi 1. Šis parametras įgauna 0 reikšmę, jei modelis neaptinka NSF, nors įmonės istorijoje jų pasitaikė, arba klaidingai indikuoja apie mokumo problemas, nors įmonė jų neturėjo. Šį atvejį atspindi 2 grafikas. Trečiajame grafike pavaizduota antrojo kriterijaus grafinė interpretacija. Tarp modelio prognozuotos ir tikrosios reikšmės yra 6 dienų laiko skirtumas, kadangi nagrinėjamu atveju modelio išėjimo slankusis langas yra 60 dienų, tai modelis apie atsirandančią NSF perspėja prieš 54 dienas.

Panaudojant šiuos du parametrus buvo įvertinta trijų modelių NSF detektavimo kokybė. Pirmiausia buvo tiriamas sprendimų medžių kolektyvo modelis, keičiant įėjimo bei išėjimo langų

trukmes tarp 30, 60 ir 90 dienų. Sprendimo medžių kolektyvo NSF prognozavimo bei signalizavimo apie ateityje atsirandančią NSF rezultatai pateikiami 7.1 lentelėje.

7.1 lentelė. Sprendimo medžių kolektyvo klasifikavimo rezultatai. NSF aptikimas ir dienos iki jų atsiradimo panaudojant skirtingų trukmių įėjimo ir išėjimo langus

		NSF aptikimas, %			Dienos iki NSF atsiradimo		
		Išėjimo langa dydis					
		30	60	90	30	60	90
Įėj.l. dydis	30	86,45	95,83	80,2	28,4	56,8	**87,7**
	60	94,79	90,62	82,29	19,23	46,4	77,1
	90	**96,87**	87,5	85,41	18,84	43,1	73,8

Kaip matyti iš prognozavimo rezultatų, naudojant 90 įėjimo bei 30 dienų išėjimo langus, pasiektas 96,87 % tikslumas, taigi vos 3 % įmonių NSF buvo nenustatytos. Minėtas modelis apie pirmą NSF signalizavo tik prieš beveik 19 dienų iki jos atsiradimo. Panaudojant 30 dienų įėjimo langą bei 90 dienų išėjimo langą, pirma NSF aptinkama vidutiniškai prieš 87,7 dienas, tačiau 20 % kompanijų, kurių istorijoje atsiranda NSF lieka nepastebėtos arba klaidingai indikuojamos kaip turinčios mokumo problemų. Naudojant 30 dienų įėjimo langą bei 60 dienų išėjimo langą pirma NSF aptinkama vidutiniškai prieš 56,8 dienas, o modelis teisingai suklasifikuoja 95,83% kompanijų. Kad būtų galima palyginti sukurto modelio rezultatus, tie patys parametrai buvo apskaičiuoti Bajeso ir atraminių vektorių mašinos modeliams. Bajeso modelio prognozavimo rezultatai pateikiami 7.2 lentelėje.

7.2 lentelė. Bajeso modelio prognozavimo rezultatai. NSF aptikimas ir dienos iki jų atsiradimo panaudojant skirtingų trukmių įėjimo ir išėjimo langus

		NSF aptikimas, %			Dienos iki NSF atsiradimo		
		Išėjimo langa dydis					
		30	60	90	30	60	90
Įėj.l. dydis	30	57,12	32,29	56,25	22	52,17	80,44
	60	64,58	61,45	66,66	21,05	55,17	**81,44**
	90	66,66	72,91	**73,95**	19,6	51,72	80,4

Bajeso modelio didžiausias klasifikavimo tikslumas (73,95%) pasiekiamas naudojant 90 dienų įėjimo bei išėjimo slankiuosius langus. Naudojant šias įėjimo duomenų charakteristikas, NSF vidutiniškai aptinkamos prieš 80,4 dienas, tuo tarpu sprendimo medžių kolektyvo modelis pirmas NSF aptinka beveik savaite vėliau (prieš 73,8 dienas), tačiau sprendimų medžio kolektyvo modelio NSF aptikimo tikslumas didesnis (85,41%).

Gauti rezultatai taip pat lyginami su klasifikavimo uždaviniams spręsti dažnai naudojamu modeliu [21] - atraminių vektorių mašina. Šio modelio prognozavimo rezultatai pateikiami 7.3 lentelėje.

7.3 lentelė. Atraminių vektorių mašinos modelio prognozavimo rezultatai. NSF aptikimas ir dienos iki jų atsiradimo panaudojant skirtingų trukmių įėjimo ir išėjimo langus

		NSF aptikimas, %			Dienos iki NSF atsiradimo		
		Išėjimo langa dydis					
		30	60	90	30	60	90
Įėj.l. dydis	30	97,91	87,5	88,54	27,97	41,1	**70,02**
	60	91,66	97,91	94,78	13,2	41,72	66,94
	90	**97,91**	93,75	92,7	11,44	39,95	67,17

Atraminių vektorių mašinos modelis beveik su visomis tiriamomis įėjimo ir išėjimo langų kombinacijomis aptiko apie 90% NSF. Geriausią rezultatą (97,91%), aptinkant NSF, atraminių vektorių mašinos modelis parodė panaudodamas 90 dienų įėjimo ir 30 dienų išėjimo langus. Nors modelis aptiko beveik visas įmones su NSF, pirma NSF buvo pastebėta tik prieš 11 dienų, tuo tarpu sprendimo medžių kolektyvo modelis šias NSF aptinka savaite anksčiau. Šiame projekte priimta, jog sistemos pagrindinis tikslas yra identifikuoti kuo daugiau įmonių, kurių istorijoje yra NSF. Taigi išrenkamas kiekvieno modelio geriausias rezultatas, tenkinantis šias sąlygas. Ištirtų modelių palyginimas pateikiamas 7.4 lentelėje.

7.4 lentelė. Modelių, tiksliausiai identifikuojančių NSF, palyginimas

Modelis	Sprendimo medžių kolektyvas	Bajeso	AVM
Įėjimo lango ilgis, d.	90	90	90
Išėjimo lango ilgis, d.	30	90	30
NSF aptikimas, %	96,87	73,95	97,91
Trukmė iki NSF atsiradimo, d.	18,84	80,4	11,72

Iš trijų ištirtų modelių geriausius rezultatus parodė atraminių vektorių mašinos modelis, kuris beveik 98% tikslumu aptiko NSF įmonėse. NSF šis modelis identifikuodavo vidutiniškai prieš 11 dienų iki jos atsiradimo. Nežymiai atsiliko sprendimo medžių kolektyvo modelis, kuris apie NSF signalizuoja prieš 18,84 dienas. NSF jis identifikavo 96,87 % įmonių, kurių istorijoje yra NSF, regresijos modelis NSF identifikavo 100 % tikslumu. Jei pageidaujama NSF identifikuoti kaip galima anksčiau ir kuo daugiau įmonių, galima remtis 7.5 lentelėje pateikiamu palyginimu.

7.5 lentelė. Modelių palyginimas, kai siekiama SF identifikuoti kaip galima anksčiau, taip pat išlaikant didžiausią NSF aptikimo procentą

Modelis	Sprendimo medžių kolektyvas	Bajeso	AVM
Įėjimo lango ilgis, d.	30	90	60
Išėjimo lango ilgis, d.	60	90	60
NSF aptikimas, %	95,83	73,95	97,91
Trukmė iki NSF atsiradimo, d.	56,8	80,4	41,72

Iš lentelės duomenų matyti, jog geriausius rezultatus rodo sprendimo medžių kolektyvo mo-

delis, kuris 95,83% kompanijų aptinka NSF ir tai daro vidutiniškai 56,8 dienos iki pirmos NSF atsiradimo. Šio modelio išėjimo ir įėjimo langai trumpiausi, lyginant su kitais dviem modeliais, taigi šis modelis reikalauja ir mažiau kompiuterio resursų, atliekant skaičiavimus. Jei sistemos tikslas NSF aptikti kuo anksčiau, neatsižvelgiant, kiek įmonių, jos identifikuojamas galima remtis 7.6 lentele.

7.6 lentelė. Modelių palyginimas, siekiant NSF identifikuoti kuo anksčiau

Modelis	Sprendimo medžių kolektyvas	Bajeso	AVM
Įėjimo lango ilgis, d.	30	60	30
Išėjimo lango ilgis, d.	90	90	90
NSF aptikimas, %	80,2	66,66	88,54
Trukmė iki NSF atsiradimo, d.	87,7	81,44	70,02

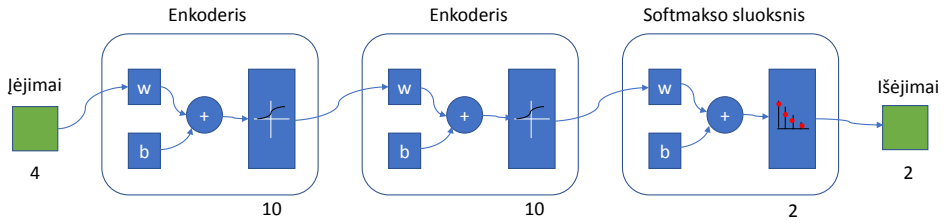
Iš lentelės duomenų matyti, jog šiuo atveju tikslinga naudoti sprendimo medžių kolektyvo modelį, kuris NSF identifikuoja prieš 87,7 dienas, tačiau tikslumas yra tik 80,2 %. Remiantis gautais rezultatais, nuspręsta įėjimo bei išėjimo langams naudoti atitinkamai 30 ir 60 dienų kuriant gilaus mokymosi modelį. Naudojant ilgesnius langus sistema apkraunama, nors ryškesnio pagerėjimo detektavimo kokybėje nėra.

8. GILAUS MOKYMOSI NEURONINIO TINKLO MODELIO KŪRIMAS

Publikuotų tyrimų sąskaitų faktūrų apmokėjimo prognozavimo srityje panaudojant gilaus mokymosi neuroninius tinklus nėra, todėl sukurtas modelis, siekiant ištirti jo veikimą, prognozuojant SF apmokėjimą.

8.1. Autoenkoderio neuroninis tinklas

Vienas iš klasifikavimo uždaviniams spręsti naudojamų gilaus mokymosi neuroninių tinklų yra autoenkoderio (angl. *autoencoder*) neuroninis tinklas [31, 32, 33]. Šis modelis bus pritaikomas mokumo indeksui prognozuoti ir realizuojamas MATLAB programavimo aplinkoje [33]. Šiame projekte bus naudojamas autoenkoderio neuroninis tinkas, susidedantis iš 3 paslėptų sluoksnių, kurių pirmieji du yra enkoderio sluoksniai, susidedantys iš 10 neuronų, bei vienas softmakso (angl. *softmax*) sluoksnis. Gilaus mokymosi neuroninio tinklo struktūra pateikiama 8.1 paveiksle.

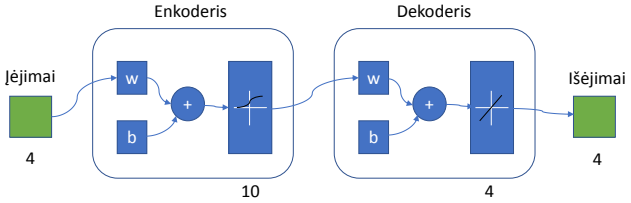


8.1 pav. Gilaus mokymosi neuroninio tinklo struktūra

Įėjimo sluoksnį sudaro anksčiau atrinkti reikšmingiausi požymiai, kurie naudojami formuojant įėjimo duomenis. Pirmiausia atliekamas pirmo enkoderio sluoksnio apmokymas, panaudojant neuroninio tinklo sluoksnio treniravimo funkciją (angl. *trainAutoencoder*):

```
autoenc1 = trainAutoencoder(X,hiddenSize,...
'ShowProgressWindow',false,...
'L2WeightRegularization',0.001,...
'SparsityRegularization',4,...
'SparsityProportion',0.05,...
'DecoderTransferFunction','purelin');
```

Šios tinklo dalies struktūra pateikiama 8.2 paveiksle.

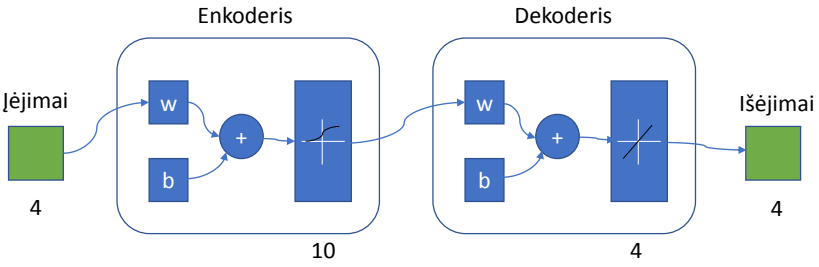


8.2 pav. Gilaus mokymosi neuroninio tinklo enkoderio sluoksnio struktūra

Šioje apmokymo dalyje modelis apsimoko ir nusistato sau svarbius požymius, neturėdamas atitinkamų išėjimo duomenų. Šie požymiai dekoduojami ir naudojami antro enkoderio sluoksniui apmokyti:

```
features1 = encode(autoenc1,X);
autoenc2 = trainAutoencoder( features1 ,hiddenSize ,...
'ShowProgressWindow',false ,...
'L2WeightRegularization' ,0.001,...
' SparsityRegularization ' ,4,...
' SparsityProportion ' ,0.05,...
'DecoderTransferFunction' ,' purelin ' ,...
'ScaleData' , false );
```

Šios dalies struktūra pateikta 8.3 paveiksle.



8.3 pav. Gilaus mokymosi neuroninio tinklo antro enkoderio sluoksnio struktūra

Nauji požymiai dar kartą dekoduojami bei naudojami softmakso sluoksnio apmokymui. Čia apmokymui jau naudojami ir išėjimo duomenys:

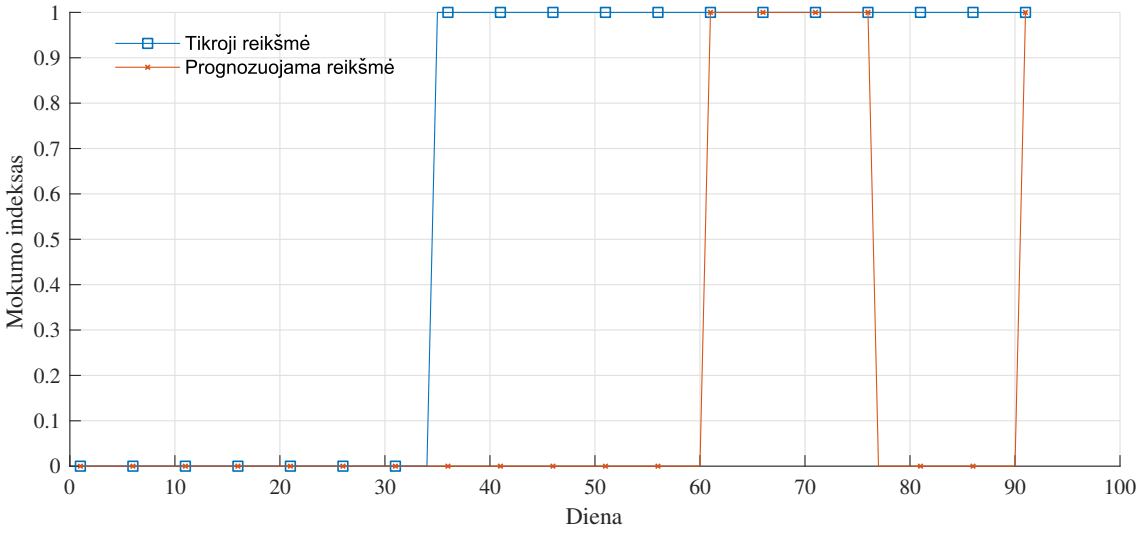
```
features2 = encode(autoenc2, features1 );
softnet = trainSoftmaxLayer( features2 ,Y,
'LossFunction' ,' crossentropy ' ,
 'ShowProgressWindow',false);
```

Apmokius šiuos tris sluoksnius jie sujungiami į vieną bei apmokomi taip pat naudojant ir išėjimo duomenis:

```
deepnet = stack(autoenc1,autoenc2, softnet )
deepnet = train (deepnet,X,Y);
```

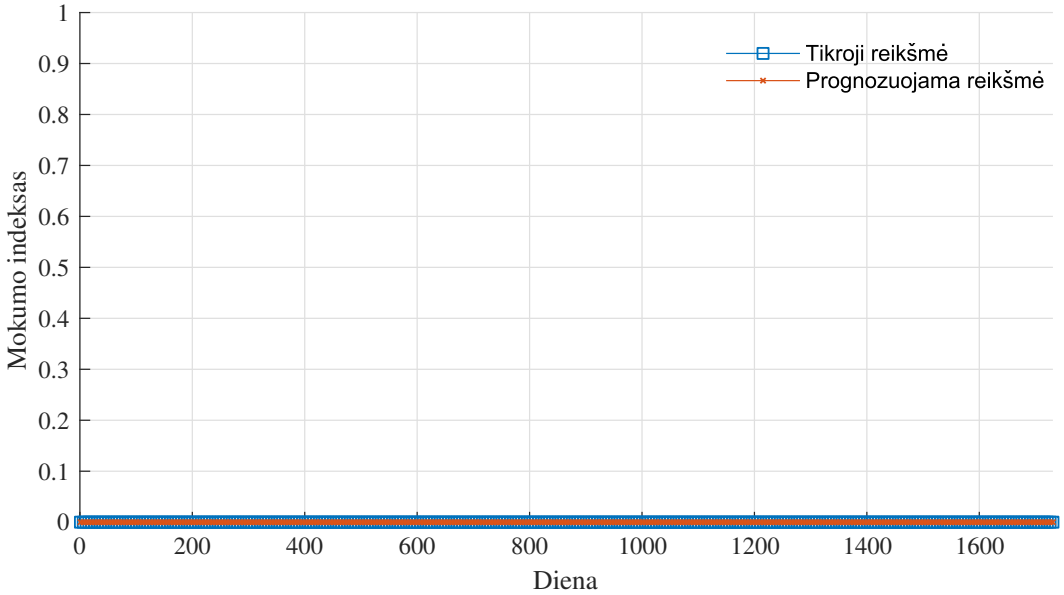
8.2. Gilaus mokymosi neuroninio tinklo detektavimo kokybės vertinimas

Apmokius gilaus mokymosi neuroninį tinklą, atliekamas šios sistemos kokybės vertinimas. Modeliui apmokyti naudotas 30-ies dienų įėjimo duomenų ir 60-ies dienų išėjimo slankusis langas. 8.4 paveiksle pateikiama mokumo indekso prognozė mažmeninės prekybos pirmajai įmonei, kaip ir 6.1 pateiktame paveiksle.



8.4 pav. Gilaus mokymosi neuroninio tinklo prognozavimo rezultatai

Gilaus mokymosi neuroninis tinklas aptiko, jog ši įmonė turės NSF, tačiau tai atliko vėliau nei sprendimų medžių kolektyvo ar atraminių vektorių mašinos modeliai. Taip pat nagrinėjamas atvejis, kai mokumo indekso prognozė atliekama įmonei, kurios istorijoje nėra NSF. 8.5 paveiksle pateikiama mokumo indekso prognozė analizuojamai įmonei.



8.5 pav. Gilaus mokymosi neuroninio tinklo prognozavimo rezultatai

Kaip ir naudojant visus kitus klasifikavimo modelius, mokumo indekso prognozė visiškai sutapo su tikrąja reikšmė ir klaidingai nemokioms priskirtų įmonių nebuvo. Sukurtas modelis taip pat buvo įvertintas remiantis 6 skyriuje aprašytais kriterijais. Šio vertinimo rezultatai pateikiami 8.1lentelėje.

8.1 lentelė. Gilaus mokymosi neuroninio tinklo NSF detektavimo kokybės vertinimas

NSF aptikimas, %	Dienos iki NSF aptikimo
82,75	49,47

Daugiau nei 80 % įmonių buvo aptiktos NSF vidutiniškai prieš 49,47 dienas iki jų atsiradimo. Gilaus mokymosi neuroninio tinklo ir kitų ištirtų modelių rezultatų palyginimas pateikiamas 8.2 lentelėje.

8.2 lentelė. Klasifikavimo modelių palyginimas

Modelis	Spr. m. kolektyvas	Bajeso	AVM	Gilaus mok. n. tinklas
NSF aptikimas, %	95,83	32,29	87,5	82,75
Trukmė iki NSF atsiradimo, d.	56,8	52,17	41,1	49,47

Ištyrus paminėtus modelius matyti, jog geriausias rezultatas pasiekiamas naudojant sprendimo medžių kolektyvo modelį. Gilaus mokymosi neuroninis tinklas taip pat galėtų būti taikomas mokumo indeksui prognozuoti, ypač jei duomenų kiekiai dideli ir reikėtų analizuoti ne 100 įmonių, bet pvz. 10000 įmonių mokumą.

IŠVADOS

1. Visose išnagrinėtose sistemose modelio įėjimo kintamuosius sudaro požymių sąrašai, apibūdinantys įmonės SF ir jų istoriją. Minėtų dviejų lygmenų naudojimas pagerina sistemos prognozavimo tikslumą. Prognozavimui naudojami modeliai: neuroniniai tinklai, atraminių vektorių mašinos modelis, atsitiktiniai miškai, logistinė regresija bei sprendimų medžiai.
2. Sukurta įmonės mokumą prognozuojanti sistema, paremta sprendimo medžių kolektyvo, atraminių vektorių mašinos, Bajeso klasifikatoriaus ir gilaus mokymosi neuroninių tinklų klasifikavimo modeliais. Klasifikavimo modeliai nepateikia klaidingų aliarmų lyginant su regresijos modeliu.
3. Sukurta reikšmingų kintamųjų atrinkimo sistema, kuri sudarydama mokymo duomenų imtį įtraukia tik tuos požymius, kurių reikšmingumas viršija optimalų slenkstį. Nustatyta, jog optimalus reikšmingumo slenkstis yra 0,35. Sistema adaptuojasi naudodama tik reikšmingiausius šešis kintamuosius modeliui apmokyti.
4. Atliktas NSF detektavimo kokybės vertinimas panaudojant du kriterijus – procentinę aptiktų kompanijų, kurių istorijoje yra NSF, išraišką bei dienų skaičių iki pirmos NSF atsiradimo signalizavus modeliui. Didėjant įėjimo slankiojo lango ilgiui, dienų skaičius iki NSF nustatymo mažėja. Remiantis banko reikalavimais, naudojant skirtingas langų kombinacijas, galima NSF identifikuoti suteikiant prioritetą arba kuo tikslesniam NSF aptikimui, arba kuo ankstesniam jų identifikavimui, bet su tam tikru klaidų skaičiumi. Norint kuo tiksliau aptikti NSF, rekomenduojama naudoti atraminių vektorių mašinos modelį bei 90 dienų įėjimo ir 30 išėjimo slankiuosius langus (NFS tikslumas – 97,91 %), o norint NSF aptikti kaip galima anksčiau, patartina naudoti sprendimo medžių kolektyvo modelį, su 30 dienų įėjimo ir 90 dienų išėjimo slankiaisiais langais (signalizuojama 87,7 dienos iki NSF atsiradimo).

LITERATŪROS ŠALTINIŲ SĄRAŠAS

1. L. Yu, S. Wang, K. K. Lai, and L. Zhou, *BioInspired Credit Risk Analysis*. Springer, 2008.
2. M. Tamari, "Financial ratios as a means of forecasting bankruptcy," *Management International Review*, pp. 15–21, 1966.
3. R. Bishnoi and R. Sahu, "A critical appraisal of company bankruptcy prediction models," *Scholedge International Journal of Business Policy & Governance ISSN 2394-3351*, vol. 2, no. 5, pp. 21–27, 2015.
4. D. Vlachos, "Neuro-fuzzy modeling in bankruptcy prediction," *Yugoslav journal of operations research*, vol. 13, no. 2, 2016.
5. J. Smirnov *et al.*, *Modelling late invoice payment times using survival analysis and random forests techniques*. PhD thesis, University of Tartu, 2016.
6. W. Hu, *Overdue invoice forecasting and data mining*. PhD thesis, Massachusetts Institute of Technology, 2016.
7. P. Hu, *Predicting and improving invoice-to-cash collection through machine learning*. PhD thesis, Massachusetts Institute of Technology, 2015.
8. S. Zeng, P. Melville, C. A. Lang, I. Boier-Martin, and C. Murphy, "Using predictive analysis to improve invoice-to-cash collection," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1043–1050, ACM, 2008.
9. M. Butkus, "Kompanijų mokumo problemų aptikimo sistemos sukūrimas ir tyrimas," bachelor's thesis, Kaunas University of Technology, Faculty of Electrical and Electronics Engineering, department of Automation and Control, 2016.
10. R. Vainienė, "Ekonomikos terminų žodynas," *Vilnius: Tyto alba*, 2005.
11. I. Justitia, "European payment report 2016". Prieiga per internetą [žiūrėta 2017.05.28]: https://www.intrum.com/globalassets/countries/norway/documents/2016/european-payment-report-europa-2016.pdf
12. D. Safavian, S. Rasoul; Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.
13. A. Liaw and M. Wiener, "Classification and regression by randomforest. r news. 2002; 2 (3): 18–22," 2016.
14. R. A. N. Freund, Yoav; Schapire, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 1, pp. 771–780, 1999.
15. V. Galvanauskas and D. Levišauskas, "Biotechnologinių procesų modeliavimas, optimizavimas ir valdymas," *Praktikumas ir uždaviniai. Vilniaus pedagoginio universiteto leidykla*, 2008.
16. V. Galvanauskas and R. Simutis, "Hibridinės procesų stebėsenos ir valdymo sistemos," *KTU leidykla „Technologija"*, 2016.
17. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, 7 2009.

18. “Matlab naudojimosi instrukcija”. Prieiga per internetą [žiūrėta 2017.05.28]: https://se.mathworks.com/help/stats/framework-for-ensemble-learning.html
19. P. Bühlmann, “Bagging, boosting and ensemble methods,” in *Handbook of Computational Statistics*, pp. 985–1022, Springer, 2012.
20. M. Bramer, *Principles of data mining*, vol. 180. Springer, 2007.
21. C.-W. Hsu, C.-C. Chang, C.-J. Lin, *et al.*, “A practical guide to support vector classification,” 2003.
22. S. Haykin and N. Network, “A comprehensive foundation,” *Neural networks*, vol. 2, no. 2004, p. 41, 2004.
23. W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.
24. P. Werbos, “Beyond regression: new fools for prediction and analysis in the behavioral sciences,” *PhD thesis, Harvard University*, 1974.
25. S. Hamori, M. Kawai, T. Kume, Y. Murakami, and C. Watanabe, “Ensemble learning or deep learning? application to default risk analysis,” *Journal of Risk and Financial Management*, vol. 11, no. 1, p. 12, 2018.
26. D. C. Ciresan, U. Meier, J. Masci, L. Maria Gambardella, and J. Schmidhuber, “Flexible, high performance convolutional neural networks for image classification,” in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, p. 1237, Barcelona, Spain, 2011.
27. H. P. Martinez, Y. Bengio, and G. N. Yannakakis, “Learning deep physiological models of affect,” *IEEE Computational Intelligence Magazine*, vol. 8, no. 2, pp. 20–33, 2013.
28. D. Ciregan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” in *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on*, pp. 3642–3649, IEEE, 2012.
29. S. Oreski, D. Oreski, and G. Oreski, “Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment,” *Expert systems with applications*, vol. 39, no. 16, pp. 12605–12617, 2012.
30. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.
31. P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.
32. J. Deng, Z. Zhang, E. Marchi, and B. Schuller, “Sparse autoencoder-based feature transfer learning for speech emotion recognition,” in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pp. 511–516, IEEE, 2013.
33. P. Li, Y. Liu, and M. Sun, “Recursive autoencoders for itg-based translation,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 567–577, 2013.

PRIEDAI

P-1. Darbo viešinimas E^2TA konferencijoje

DIPLOMA

Mantas Butkus

Prepared and presented a paper in the conference

E²TA - 2017

On the topic:

Kompanijų mokumo problemų aptikimo sistemos kūrimas ir tyrimas

Prof. Algimantas Valinevičius
Chairman of the organizing committee
Dean of the Electronics and Electrical Engineering Faculty

Kaunas, 2017 05 11