

Potencialių pardavimų baigties prognozavimas naudojant mašininio mokymosi modelius

Eimantė Žemytė

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Naugarduko g. 24, LT-03225 Vilnius
eimante.zemyte@mif.stud.vu.lt

Santrauka. Potencialių pardavimų baigties prognozavimas atliktas naudojant klasifikacinius: logistinės regresijos ir atsitiktinio miško (angl. *Random Forest*) modelius. Duomenų pagrindinis šaltinis – klientų ryšių valdymo sistema „Microsoft Dynamics 365 CRM“. Papildomi šaltiniai – „Sodros“ ir „Registru centro“ viešųjų duomenų rinkiniai. Kryžminės validacijos metu nustatyta, kad atsitiktinio miško modelis yra veiksmingesnis (vidutinis tikslumas 91 %), palyginti su logistinės regresijos modeliu (vidutinis tikslumas 74 %). Taip pat nustatyta, kad pardavimų baigties prognozei, remiantis Gini indeksu, statistiškai reikšmingiausi „Microsoft Dynamics 365 CRM“ duomenys.

Raktiniai žodžiai: klientų valdymo sistema (angl. *Customer Relationship Management, CRM*), verslas verslui (angl. *Business To Business, B2B*), logistinė regresija, atsitiktinis miškas (angl. *Random Forest*)

Įvadas

Tikslus pardavimų baigties prognozavimas yra svarbus verslams, nes remdamasi prognozės rezultatais įmonė gali planuoti produkciją, žmogiškuosius resursus ir inventorių [9]. Didžiausias iššūkis įmonėms – sukurti patikimą prognozės įrankį. Įmonės gali nesiryžti investuoti į matematinius skaičiavimus grįstus prognozės modelius, bet rinktis paprastesnį būdą – pasitikėti vadybininko ekspertine nuomone. Būtent čia įmonės susidūria su esmine problema – vadybininko ekspertinė nuomonė yra subjektyvi ir vertinimai gali smarkiai skirtis nuo galutinių pardavimų rezultatų. Net garsiausių pasaulio įmonių vadovai ir atstovai yra apsirikę teikdami prognozes. Pavyzdžiui, buvęs „Microsoft“ generalinis direktorius S. Balmeris 2007 metais teigė, kad „iPhone“ niekada neįgis reikšmingos rinkos dalies. Be to, kitos pasaulio milžinės „IBM“ atstovai 1959 metais įmonei „Xerox“, užsiimančiai

spausdintuvų prekyba, prognozavo, kad pasaulio rinkoje bus įmanoma parduoti daugiausiai vos penkis tūkstančius spausdintuvų [10]. Akivaizdu, jog prognozės nepasitvirtino: „Apple“ užima 27.78 % viso pasaulio rinkos [6], o „Xerox“ 2024 m. sugeneravo virš 6 mlrd. JAV dolerių pajamų [7]. Dėl šios priežasties svarbu sukurti pardavimų modelį, grįstą ne vadybininko intuicija, tačiau objektyviais duomenimis.

Klientų ryšių valdymo sistemos (angl. *Customer Relationship Management*, CRM) tapo nepakeičiamu įrankiu šiuolaikinių verslų pardavimų srityje. Jų duomenys yra vieni iš svarbiausių visų įmonės valdomų duomenų [3]. Sistemoje galima rasti komunikacijos su potencialiais klientais istoriją, pasiūlymų detales ir kitas metrikas, leidžiančias įmonėms analizuoti klientų eigles ir optimizuoti pardavimų strategijas. Klientų ryšių valdymo sistemoje „Microsoft Dynamics 365 CRM“ pardavimai yra išskaidyti į penkis pardavimo etapus [4]:

1. Potencialių klientų (angl. *leads*) pritraukimas. Identifikuojami ir pritraukiami galimi pirkėjai. Šio etapo metu vyksta pirmieji skambučiai (kitaip vadinami *šaltaisiais skambučiais*) ir nustatomi kliento patiriami iššūkiai.
2. Jeigu įmanoma pasiūlyti tinkamą produktą ar paslaugą kliento sunkumams išspręsti, šie sprendimai sistemoje suformuojami kaip „galimybės“ (angl. *opportunities*) – potencialūs paslaugos pardavimai, atitinkantys kliento poreikius. Šio etapo metu vyksta papildomi pokalbiai siekiant suderinti tinkamiausią, personalizuotą sprendimą ir optimalią kainą.
3. Suderinus abiems pusėms tinkamas sąlygas, klientui pateikiamas komercinis pasiūlymas.
4. Patvirtinamas užsakymas ir sudaroma sutartis. Jei sutartis pasirašoma, galimybei (potencialus pardavimas) suteikiamas statusas „laimėta“, nes pardavimas įvyko.
5. Suteikus paslaugą/produktą, parengiama ir išrašoma sąskaita už atliktas paslaugas.

Šio tyrimo **tikslas** yra sukurti kuo tikslesnį mašininio mokymosi modelį, skirtą prognozuoti, ar konkretus potencialus pardavimas virs užsakyму, eliminuojant vadybininko ekspertinės nuomonės tiesioginės įtakos prognozavimo rezultatams veiksnį. Pasirinkta prognozuoti potencialių pardavimų baigtį, nes agreguoti prognozių rezultatai leidžia įsivertinti potencialias par-

davimų pajamas. Pavyzdžiui, nustačius, kad prognozės neatitinka lūkesčių, galima didinti pajėgas pritraukiant potencialius klientus, kad atsirastų daugiau potencialių pardavimų ir būtų didesnė tikimybė pasiekti įmonės finansinius tikslus. Be to, atsižvelgiant į prognozės rezultatus galima racionaliai planuoti žmogiškuosius išteklius ir finansus [2].

Išsikelti darbo uždaviniai:

1. Paruošti realius IT sektoriaus įmonės duomenis iš „Microsoft Dynamics 365 CRM“ platformos.
2. Surasti ir papildyti duomenų rinkinį viešai prieinamais duomenimis apie kiekvieną potencialų klientą (įmonę).
3. Nustatyti, kurie rodikliai yra reikšmingiausi prognozuojant pardavimus.
4. Sukurti ir palyginti logistinės regresijos ir atsitiktinio miško (angl. *Random forest*) modelius, skirtus potencialių pardavimų sėkmės prognozėms gauti.

1 Duomenys ir metodai

Šio empirinio tyrimo metu yra naudojami tikri informacinių technologijų sektoriaus įmonės duomenys, gauti iš klientų ryšių valdymo sistemos „Microsoft Dynamics 365 CRM“. Be to, šio tyrimo metu naudojami ir viešai prieinami valstybiniai duomenys: „Sodros ir „Registų centro“ –atvirų, t. y. viešųjų duomenų, rinkiniai.

Galutinį duomenų masę sudaro šie elementai:

- Klientų potencialių pardavimų istorija:
 - ankstesnių sėkmingų ir nesėkmingų pardavimų baigtis;
 - įgyvendintų projektų punktualumas (pavėluotai įgyvendintų projektų dalis);
 - atsakingo pardavimų vadybininko darbo patirtis (sėkmingų pardavimų dalis ir darbo stažas įmonėje);
 - kliento (įmonės) identifikavimo informacija (juridinio asmens, PVM mokėtojų kodai).
- Viešai prieinami valstybiniai duomenys (prijungiami prie klientų naudojant juridinio asmens, PVM mokėtojų kodus):
 - „Sodros“ duomenys: informacija apie apdraustųjų skaičių ir vidutinį darbo užmokestį kiekvienam klientui;
 - „Registų centro“ duomenys: juridinių asmenų finansinės atskaitomybės dokumentai, įskaitant ekonominės veiklos rūšį, pelno (nuostolio) ataskaitas ir balansus.

Galutinė imtis – 2187 potencialių pardavimų (galimybių) duomenys, kurių baigtis gali būti sėkminga arba nesėkminga (pardavimas neįvyko). Bendras duomenų rinkinio dydis – 455 įmonės, iš kurių 141 įmonė turėjo tik vieną pardavimo galimybę (t. y., šie klientai pirmą kartą kreipėsi į įmonę jau su konkrečiu poreikiu dėl specifinio sprendimo – paslaugos ar produkto). Iš viso buvo nagrinėjama 20 požymių (angl. *features*).

Atsižvelgiant į tai, kad įmonių finansinių ataskaitų informacijos detalumas priklauso nuo įmonės specifikos (pvz., įmonės dydžio) [1], papildomai apskaičiuoti keli pagrindiniai finansiniai rodikliai [6]:

- Nuosavo kapitalo pelningumas (grąža) (angl. trumpinys *ROE*):

$$ROE = \frac{\text{grynasis pelnas}}{\text{nuosavas kapitalas}}$$

Šis rodiklis suteikia informaciją apie įmonės praėjusio laikotarpio investicijų į įmonę grąžą;

- Turto pelningumas (grąža) prieš mokesčius (angl. trumpinys *ROA*):

$$ROA = \frac{\text{grynasis pelnas}}{\text{vidutinis turtas}}$$

Šis rodiklis padeda įvertinti, kaip efektyviai įmonė panaudoja turtą;

- Veiklos pelningumas prieš palūkanas ir mokesčius (angl. trumpinys *EBIT*):

$$EBIT = \frac{\text{pelnas (prieš apmokestinimą)}}{\text{pajamos}}$$

Šis rodiklis padeda prognozuoti, kokią pelną uždirbtų įmonė, neįgydama finansinių skolų;

- Skolos ir nuosavo kapitalo santykis (angl. trumpinys *D/E*):

$$D/E = \frac{\text{mokėtinos sumos ir įsipareigojimai}}{\text{nuosavas kapitalas}}$$

Vienas iš pagrindinių rodiklių, indikuojančių finansinę riziką.

Taigi visi šie pagrindiniai rodikliai suteikia galimybę apžvelgti bendrą potencialių klientų finansinę situaciją: ar klientas mokus, ar turi reinvestavimo potencialo.

Galimybės konversijos į užsakymą prognozuoti naudojami du klasifikaaciniai modeliai: logistinė regresija ir atsitiktinis miškas (angl. *Random forest*). Šių modelių pasirinkimas suteikia galimybę efektyviai įvertinti potencialaus

pardavimo baigtį, nes jų rezultatas pateikia binarinį atsakymą – sėkmę arba nesėkmę. Logistinės regresijos modelis vertina ryšį tarp nepriklausomų kintamųjų ir kategorinio priklausomo kintamojo ir prognozuoja tikimybę, ar potencialus pardavimas priklauso „laimėtų pardavimų“ kategorijai. Šis modelis pasirinktas dėl palyginti nesudėtingo įgyvendinimo ir aiškiai interpretuojamų koeficientų, suteikiančių galimybę identifikuoti reikšmingiausius požymius, darančius įtaką prognozės rezultatui. Visgi minėtinas ir tam tikras šio modelio trūkumas: kadangi logistinė regresija yra jautri multikolinearumui, tai gali sukelti nestabilius koeficientų įverčius ir sumažinti modelio patikimumą. Dėl šios priežasties lyginamosios analizės tikslu taip pat pasirinktas atsitiktinio miško metodas, kuris yra atsparesnis multikolinearumui, nes vietoj linijinių kombinacijų naudojami sprendimų medžiai – skaidomi pavieniai kintamieji kiekviename mazge, taip sumažinant kintamųjų tarpusavio priklausomybės įtaką modeliui [11].

Abiems modeliams atliekama kryžminė validacija (10 grupių) ir pagal gautas testinės paklaidos reikšmes (angl. *test error values*) išrenkamas efektyvesnis modelis. Taikant logistinės regresijos modelį, užtikrinama, jog kintamieji yra statistiškai reikšmingi (reikšmingumo lygmuo $p < 0,05$) ir nėra multikolinearumo tarp kintamųjų ($VIF < 4$). Atsitiktinio miško modelio hiperparametrai parinkti naudojant 50 iteracijų atsitiktinės paieškos kryžminę validaciją (angl. *Randomized Search CV*) su kryžmine validacija (10 kartų). Galutinis modelis parinktas pagal geriausią ROC AUC rezultatą ir OOB įvertį (angl. *out-of-bag score*).

Galutiniame atsitiktinio miško modelyje naudoti hiperparametrai:

- medžių skaičius ($n_estimators$) = 500;
- maksimalus medžių gylis (max_depth) = 30;
- maksimalus išsišakojančių mazgų (angl. *child nodes*) skaičius ($max_features$) = kvadratinė šaknis ($\sqrt{p}$);
- minimali imtis, reikalinga vidinio mazgo (angl. *internal node*) padalijimui ($min_samples_split$) = 2;
- klasių svoris ($class_weight$) = balansuotas.

2 Rezultatai

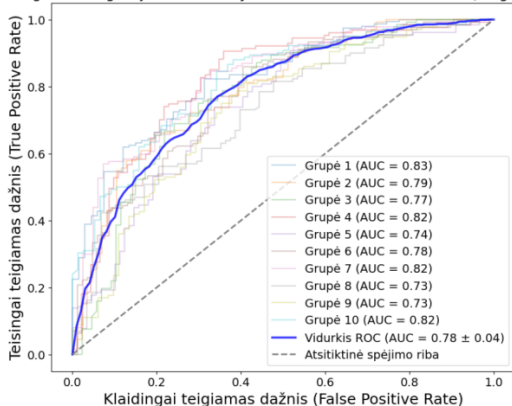
Atsitiktinio miško (angl. *Random forest*) modelis veiksmingesnis už logistinės regresijos modelį. Pagal ROC kreives (žr. 1 pav. ir 2 pav.), vaizduojančias jautrumo ir specifiškumo sąryšį, matoma, kad plotas po ROC kreive (AUC)

yra didesnis atsitiktinio miško modeliui (čia vidutinis AUC logistinei regresijai – 0,78; atsitiktiniam miškui – 0,97). Atsitiktinis miškas itin optimaliai klasifikuoja galimybių baigtį;

Kryžminės validacijos metu gautas vidutinis tikslumas:

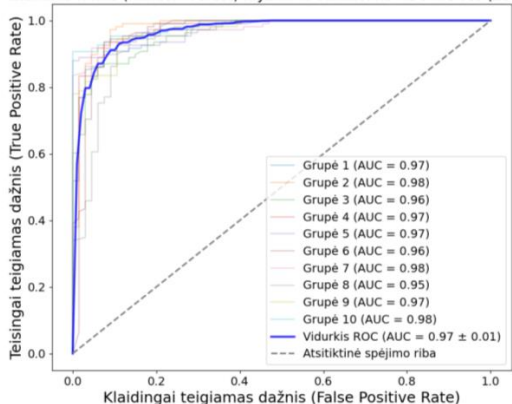
- logistinės regresijos – 74 % (0,740 ($\pm 0,032$))
- atsitiktinio miško – 91 % (0,913 ($\pm 0,020$))

Logistinės regresijos modelio kryžminio validavimo ROC kreivės (10 grupių)



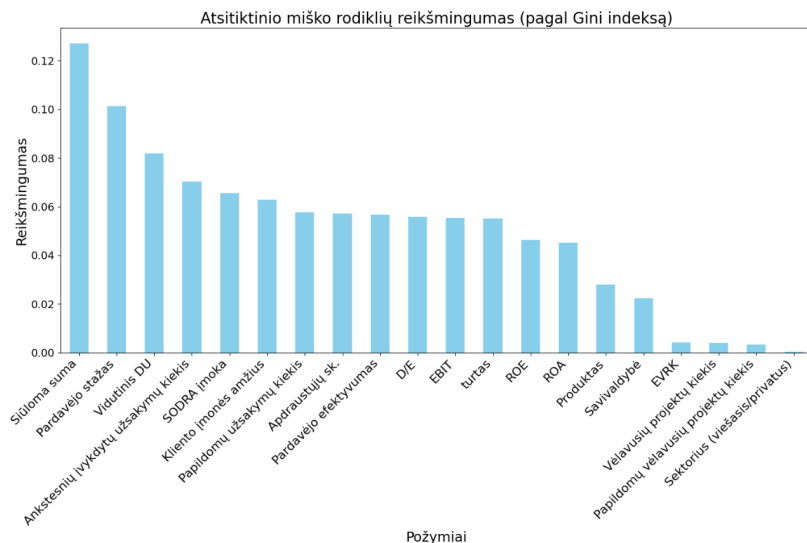
1 pav. Logistinės regresijos ROC kreivė

Atsitiktinio miško (Random Forest) kryžminio validavimo ROC kreivės (10 grupių)



2 pav. Atsitiktinio miško ROC kreivė

Kadangi atsitiktinio miško modelis gautas tikslesnis, jame nustatomas kintamųjų reikšmingumas naudojant *Gini indeksą* [5]:



3 pav. Atsitiktinio miško modelio rodiklių reikšmingumas pagal Gini indeksą

Reikšmingiausi atsitiktinio miško rodikliai (žr. 3 pav.): derybose pasiūlyta preliminarinė suma, pardavėjo stažas (dienomis), vidutinis potencialaus kliento (įmonės) darbo užmokestis, kliento ankstesnių užsakymų kiekis. Mažiausiai reikšmingi (*Gini indeksas* < 0,01): įmonės sektorius – viešasis/privatus, kliento anksčiau įvykdytų, bet vėlavusių užsakymų kiekis, ekonominės veiklos rūšis (*EVVK*). Reikšmingiausi finansiniai rodikliai: skolos ir nuosavo kapitalo santykis (*D/E*), veiklos pelningumas prieš palūkanas ir mokesčius (*EBIT*).

Iš gautų duomenų analizės rezultatų matoma, kad be agreguotų istorinių galimybių duomenų, statistiškai reikšmingi yra ir atvirų duomenų šaltiniai. Pastebėta, kad projektų vėlavimai, ekonominės veiklos rūšis bei juridinio asmens sektoriaus tipas daro nedidelę įtaką prognozei. Verta paminėti, kad dauguma įmonės klientų priklauso privačiam sektoriui ir jų ekonominė veikla sutampa. Išskirtos vos dvi skirtingos ekonominės veiklos rūšies kategorijos. Taigi visi mažiausiai reikšmingi kintamieji (kurių *Gini indeksas* < 0,01) pasižymi maža grupių įvairove (vos 2 unikalios reikšmės imtyje).

Išvados

Empirinio tyrimo metu įvertinus ir palyginus potencialių pardavimų baigties prognozavimo tikslu naudojamus klasifikacinius logistinės regresijos ir atsitiktinio miško (angl. Random Forest) modelius nustatyta, kad atsitiktinio miško modelis yra veiksmingesnis, palyginti su logistinės regresijos modeliu, nes jis 17 % tiksliau prognozuoja, kiek įmonės turimų galimybių (angl. *opportunities*) virs užsakymais.

Pagal *Gini indeksą* nustatyta, kad didžiausią įtaką modeliui daro „Microsoft Dynamics 365 CRM“ platformoje vedami duomenys. Itin svarbų vaidmenį atlieka ir nemokamai, viešai prieinami „Sodros“ duomenys ir „Registų centro“ prieinamos finansinės ataskaitos (iš jų išvesti finansiniai rodikliai).

Pabrėžtina, jog iš finansinių rodiklių aktualiausi yra skolos ir nuosavo kapitalo santykis (D/E) ir EBIT. Tad apibendrinus empirinio tyrimo rezultatus galima daryti prielaidą, jog pardavimų prognozei aktualiausia su skolomis ir potencialiu reinvestavimu susijusi informacija.

Literatūra

- [1] Arkan, T. (2016). The importance of financial ratios in predicting stock price trends: A case study in emerging markets. *Zeszyty Naukowe Uniwersytetu Szczecińskiego Finanse Rynki Finansowe Ubezpieczenia*, 1, 13–26.
- [2] Berko, A., Pelek, I., & Hlova, P. (2024). Analysis and Formation of Sales Forecasts in CRM Systems. *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems (CoLInS 2024), Volume III: Intelligent Systems Workshop*, 288–297.
- [3] Hallikainen, H., Savimäki, E., & Laukkanen, T. (2020). Fostering B2B sales with customer big data analytics. *Industrial Marketing Management*, 86, 90–98.
- [4] Microsoft. (dokumentacija). Nurture sales from lead to order <https://learn.microsoft.com/It-It/dynamics365/sales/nurture-sales-from-lead-order-sales>
- [5] Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random Forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review. *Frontiers in Aging Neuroscience*, 9, 329.
- [6] StatCounter (2025). Mobile Vendor Market Share. <https://gs.statcounter.com/vendor-market-share/mobile> (pasiekta 2025-03-27)
- [7] Yahoo Finance (2025). Xerox Corporation (XRJ) Financials. <https://finance.yahoo.com/quote/XRJ/financials/> (pasiekta 2025-03-27)
- [8] Vertybinių popierių birža NASDAQ OMX Vilnius (2010). *Įmonių finansinė analizė*. Rodiklių skaičiavimo metodika
- [9] Vivian, M. E., & Regi, R. G. (2019). *Exploring the suitability of support vector regression and radial basis function approximation to forecast sales of Fortune 500 companies*. *Advances in Business and Management Forecasting*, 13, 3–5.
- [10] Schneider, J. V., Alavi, S., Guba, J. G., Wieseke, J., & Schmitz, Ch. (2021). When do forecasts fail and when not? Contingencies affecting the accuracy of sales managers' forecast regarding the future business situation. *Journal of Personal Selling & Sales Management*, 41(3), 218–232.
- [11] James, G., Witten, D., Hastie, T., Tibshirani, R. ir Taylor, J. (2023). *An introduction to statistical learning: Python edition*. Springer.