Pokročilé modelovanie rizika v povinnom zmluvnom poistení: Využitie štatistických metód a strojového učenia

Advanced risk modeling in motor third party liability insurance: Using statistical methods and machine learning

Peter Schmidt[1], Silvia Zelinová[2], Eva Rakovská[3]

Abstrakt

Cieľom článku je analyzovať možnosti kombinácie tradičných štatistických metód a modelov strojového učenia (ML) pri spracovaní údajov v oblasti povinného zmluvného poistenia (PZP). Východiskom je použitie generalizovaných lineárnych modelov a kontrastnej analýzy na štatistickú segmentáciu poistencov podľa rizikových faktorov, ako sú vek vodiča, značka a výkon vozidla či lokalita. V ďalšej fáze boli implementované ML algoritmy, najmä Random Forest, na imputáciu chýbajúcich údajov a predikciu závažnosti poistných udalostí (claim severity). Výsledky ukazujú, že strojové učenie dosahuje vyššiu predikčnú presnosť než tradičné modely, pričom zároveň identifikujú kľúčové faktory rizika. Kombinovaný analytický prístup spája výhody oboch metodík – vysvetliteľnosť a regulačnú transparentnosť štatistiky so škálovateľnosťou a výpočtovým výkonom ML. Záverom štúdia odporúča hybridný modelovací rámec ako optimálny nástroj pre segmentáciu poistencov, dynamickú cenotvorbu a efektívne riadenie rizika v PZP.

Kľúčové slová

PZP, poistné riziko, ML, predikcia poistného plnenia, Random Forest

Abstract

This article aims to analyze the potential of combining traditional statistical methods with machine learning (ML) models in the context of compulsory motor third party liability insurance (MTPL). The study begins with the application of generalized linear models and contrast analysis to statistically segment policyholders based on risk factors such as driver age, vehicle brand and engine power, or location. In the next phase, ML algorithms - particularly Random Forest - are used for imputing missing data and predicting claim severity. The results show that machine learning achieves higher predictive accuracy ($R^2 = 0.76$) than traditional models ($R^2 = 0.71$), identifying key risk drivers. The combined analytical approach leverages the strengths of both methodologies—interpretability and regulatory transparency of statistics, alongside the scalability and computational power of ML. The study concludes by

[1] Ekonomická univerzita v Bratislave, Fakulta hospodárskej informatiky, Katedra aplikovanej informatiky, Dolnozemská cesta 1, 852 35, Bratislava, Slovensko, peter.schmidt@euba.sk (*korešpondujúci autor.

[2] Ekonomická univerzita v Bratislave, Fakulta hospodárskej informatiky Katedra matematiky a aktuárstva, Dolnozemská cesta 1, 852 35, Bratislava, Slovensko silvia.zelinova@euba.sk.

[3] Ekonomická univerzita v Bratislave, Fakulta hospodárskej informatiky, Katedra aplikovanej informatiky, Dolnozemská cesta 1, 852 35, Bratislava, Slovensko, eva.rakovska@euba.sk.

recommending a hybrid modeling framework as an optimal tool for policyholder segmentation, dynamic pricing, and effective risk management in MTPL.

Key words

motor third party liability insurance, insurance risk, machine learning, claim prediction, Random Forest

JEL classification

C51, C55, G22, C38

1 Úvod

Motorové vozidlo je dôležitým prvkom mobility v moderných spoločnostiach, no jeho prevádzka so sebou prináša riziko dopravných nehôd a následných finančných škôd. Povinné zmluvné poistenie (PZP) je preto kľúčovým pilierom poistného systému v mnohých krajinách vrátane Slovenska. Tento produkt zabezpečuje krytie škôd spôsobených prevádzkou motorového vozidla tretím stranám, čím chráni nielen vodičov, ale aj ostatných účastníkov cestnej premávky. Vzhľadom na jeho povinný charakter je efektívne a presné spracovanie veľkého množstva poistných údajov nevyhnutnosťou pre poisťovne pri riadení rizík, segmentácii poistencov a optimalizácii poistných produktov.

Tradične sa v oblasti PZP využívajú aktuárske a štatistické metódy na výpočet poistného, pričom zovšeobecnené lineárne modely (GLM) sú jedným z hlavných nástrojov na stanovenie rizikových profilov poistencov (Šoltés et al., 2019). V posledných rokoch však v poisťovníctve získavajú čoraz väčšiu popularitu modely strojového učenia (ML), ktoré umožňujú efektívnejšie spracovanie a analýzu rozsiahlych datasetov, predikciu poistných udalostí, či dokonca imputáciu chýbajúcich údajov (Richman, 2018). Spojenie tradičných štatistických metód s ML modelmi preto predstavuje inovatívny prístup k segmentácii poistencov, cenotvorbe a zlepšeniu kvality spracovania poistných dát (Hanafy et al., 2021).

Hlavným cieľom tohto článku je analyzovať možnosti kombinácie štatistických a ML metód pri spracovaní dát v oblasti PZP. V prvej časti sa zameriame na využitie kontrastnej analýzy pri segmentácii poistencov a odhade závažnosti poistných udalostí. Následne predstavíme postupy implementácie ML modelov na spracovanie poistných údajov, predovšetkým metódu Random Forest na imputáciu chýbajúcich hodnôt a predikciu poistného rizika (Rawat et al., 2021). Výsledkom bude porovnanie efektívnosti týchto prístupov a diskusia o možnostiach ich kombinácie v poistnej praxi.

2 Teoretický rámec

Spracovanie dát v povinnom zmluvnom poistení (PZP) je komplexný proces, ktorý si vyžaduje kombináciu aktuárskych a štatistických metód spolu s modernými prístupmi založenými na strojovom učení (ML). Cieľom tohto rámca je vysvetliť základné teoretické východiská oboch prístupov a ukázať ich potenciálne prepojenie pri analýze rizika, segmentácii poistencov a určovaní výšky poistného.

2.1 Aktuárske a štatistické metódy v segmentácii poistencov

V tradičných prístupoch k výpočtu poistného sa využívajú modely z oblasti aktuárskych vied a štatistiky. Zovšeobecnené lineárne modely (GLM) patria medzi najčastejšie používané

metódy v neživotnom poistení, pretože umožňujú efektívne modelovanie pravdepodobnosti poistných udalostí a ich závažnosti (Reiff et al., 2022). Pri výpočte poistného sa zohľadňujú viaceré premenné, ako napríklad:

- Technické parametre vozidla (výkon, hmotnosť, značka, rok výroby),
- Charakteristiky vodiča (vek, lokalita, história škodovosti),
- Účel využitia vozidla (osobné, firemné, taxislužba).

Kontrastná analýza v segmentácii poistencov

Jednou z menej využívaných, no veľmi efektívnych štatistických metód v poisťovníctve je kontrastná analýza. Tá umožňuje podrobnejšie skúmať rozdiely medzi skupinami poistencov a identifikovať relevantné faktory, ktoré najviac ovplyvňujú výšku škôd. Na rozdiel od tradičných priemerných hodnôt, ktoré môžu byť skreslené v dôsledku nerovnomerného rozdelenia údajov, kontrastná analýza využíva marginalizované priemery (Least Squares Means – LS-means) na elimináciu vplyvu ostatných faktorov.
V oblasti PZP umožňuje kontrastná analýza:

- Presnejšie rozdelenie poistencov do homogénnych skupín, kde v rámci segmentu neexistujú významné rozdiely v škodovosti;
- Odhalenie interakčných efektov medzi premennými, napríklad vplyv veku vodiča na závažnosť škôd v závislosti od značky vozidla;
- Lepšie modelovanie cenotvorby poistného, keďže poisťovne môžu stanoviť spravodlivejšie sadzby na základe štatisticky významných rozdielov medzi skupinami vodičov.

V tradičných aktuárskych modeloch sa však často zanedbáva komplexnosť vzťahov medzi jednotlivými faktormi, čo môže viesť k suboptimálnym výsledkom. Strojové učenie zohráva kľúčovú úlohu práve v tejto fáze analytického procesu, keďže umožňuje identifikáciu latentných vzorov v dátach a zároveň poskytuje nástroje na efektívne spracovanie rozsiahlych dátových súborov (Baudry & Robert, 2019).

2.2 Strojové učenie a jeho využitie v analýze PZP

Modely strojového učenia (ML) umožňujú poisťovniam nielen presnejšie predikovať riziká, ale aj efektívnejšie spracovávať veľké objemy dát. V oblasti PZP sa ML využíva predovšetkým na:

- Imputáciu chýbajúcich hodnôt – pomocou metód ako Random Forest Imputation je možné doplniť chýbajúce údaje v databázach bez skreslenia výsledkov;
- Predikciu pravdepodobnosti poistných udalostí – ML algoritmy dokážu identifikovať vodičov s vyšším rizikom nehody na základe historických údajov;
- Predikcia poistného na základe parametrov a škodovosti klienta;
- Odhalenie anomálií a podvodov – pokročilé algoritmy ako neurónové siete alebo autoenkódery dokážu detegovať podozrivé poistné udalosti.

Random Forest a jeho využitie pri spracovaní PZP dát

Jeden z najpoužívanejších algoritmov v poisťovníctve je Random Forest, ktorý patrí medzi metódy súborového učenia (ensemble learning). Tento algoritmus využíva viacero rozhodovacích stromov na vytvorenie robustného prediktívneho modelu. Jeho hlavné výhody zahŕňajú:

- Odolnosť voči pretrénovaniu – vďaka bootstrap agregácii (bagging) je model menej náchylný na chyby spôsobené šumom v dátach;

- Flexibilitu pri spracovaní rôznych typov údajov – dokáže pracovať s číselnými aj kategorizovanými premennými;
- Možnosť identifikácie kľúčových faktorov – pomocou metódy feature importance je možné určiť, ktoré premenné najviac ovplyvňujú škodovosť.

V poisťovníctve sa Random Forest využíva na imputáciu chýbajúcich hodnôt, kde predikuje chýbajúce údaje na základe podobnosti s existujúcimi poistnými zmluvami. Tento prístup umožňuje poisťovniam zachovať úplnosť databáz a minimalizovať chyby spôsobené nekompletnými údajmi.

2.3 Synergia tradičných aktuárskych metód a strojového učenia

Aj keď ML modely ponúkajú mnoho výhod, nie sú dokonalým riešením (Grize et al., 2020). Štatistické metódy, ako GLM a kontrastná analýza, poskytujú interpretovateľnosť výsledkov a overiteľnosť modelov, čo je dôležité v regulačnom prostredí poisťovníctva. Na druhej strane, ML modely dokážu efektívne spracovať nelineárne vzťahy a pracovať s rozsiahlymi datasetmi. Preto kombinácia oboch prístupov:

- Štatistické modely (GLM + kontrastná analýza) → Zabezpečujú teoretickú správnosť, vysvetliteľnosť a konzistentnosť výsledkov;
- ML modely (Random Forest, neurónové siete) → Pomáhajú odhaliť skryté vzory a optimalizovať procesy spracovania dát;

predstavuje optimálne riešenie.

Táto kombinácia umožňuje poisťovniam efektívne predikovať riziká, minimalizovať chyby v údajoch a zabezpečiť spravodlivejšie oceňovanie poistného.

3 Metodika

Táto časť načrtáva metodický prístup použitý na analýzu údajov v kontexte povinného zmluvného poistenia zodpovednosti za škodu spôsobenú prevádzkou motorového vozidla. Dôraz sa kladie na súbor údajov, techniky predbežného spracovania údajov, stratégie štatistického modelovania a implementáciu algoritmov strojového učenia. Cieľom je poskytnúť transparentný a replikovateľný rámec na analýzu rozsiahlych poistných údajov pomocou tradičných aj moderných analytických nástrojov.

3.1 Opis datasetu

Empirická analýza je založená na vlastnom súbore údajov poskytnutom nemenovanou poisťovňou, ktorý pokrýva poistné zmluvy a záznamy o škodách počas deviatich rokov (2016–2024). Súbor údajov obsahuje viac ako 540 000 poistných zmlúv a približne 33 500 poistných udalostí súvisiacich s PZP, čo ponúka bohatý základ pre štatistické modelovanie a aplikácie strojového učenia. Všetky údaje v dátovom súbore boli anonymizované.

Z úplnej databázy bol extrahovaný výber relevantných premenných a kategorizovaný do troch hlavných skupín:

- *Technické parametre vozidla:*
  - Výkon motora (kW),
  - Hmotnosť vozidla (kg),
  - Rok výroby,
  - Značka vozidla,
- *Charakteristiky poistníka:*
  - Vek vodiča,
  - Lokalita (okres trvalého bydliska),

  - História škodovosti (bonus/malus),
- *Poistné a poistné udalosti:*
  - Výška ročného poistného (€),
  - Počet poistných udalostí,
  - Výška škôd (claim severity – CS).

Pred analýzou prešiel súbor údajov komplexným procesom čistenia. Všetky záznamy s neúplnými alebo neobnoviteľnými informáciami boli vylúčené. Okrem toho sa vykonala séria kontrol konzistencie, aby sa zabezpečila spoľahlivosť údajov. To zahŕňalo overenie číselných formátov, odstránenie duplicitných záznamov a identifikáciu odľahlých hodnôt alebo anomálií, ktoré by mohli naznačovať chyby pri zadávaní údajov alebo štrukturálne nezrovnalosti. Takáto príprava údajov je nevyhnutná pre štatistickú integritu aj robustnosť modelov strojového učenia. Zabezpečuje, že následné analýzy nie sú ovplyvnené chybnými záznamami a že modely sú trénované na kvalitných reprezentatívnych údajoch.

3.2 Predspracovanie dát

Efektívne predspracovanie údajov je kritickým krokom pri vytváraní robustných štatistických modelov a modelov strojového učenia (James et al., 2023). Zabezpečuje kvalitu údajov, rieši nezrovnalosti a pripravuje súbor údajov na spoľahlivú a nezaujatú analýzu. Pracovný postup predbežného spracovania pre túto štúdiu zahŕňal niekoľko kľúčových fáz (Provost, 2023).

Čistenie a kontrola kvality dát

Počiatočná fáza zahŕňala komplexnú kontrolu súboru údajov s cieľom identifikovať a opraviť problémy s kvalitou údajov. Riešilo sa niekoľko bežných problémov:

Identifikácia chýbajúcich hodnôt – Najčastejšie medzery boli pozorované pri premenných ako výkon motora, hmotnosť vozidla, bydlisko a rok narodenia.

Odstránenie extrémnych alebo nepravdepodobných hodnôt – Boli vylúčené záznamy s nereálnymi údajmi, ako sú uvedené vozidlá s výkonom motora 0 kW alebo vek vodiča pod 18 alebo nad 100 rokov.

Korekcia dátových formátov – bola vykonaná štandardizácia pre číselné polia (napr. oddeľovače desatinných miest) a kódovanie znakov, aby sa zabezpečila konzistentnosť v rámci celého súboru údajov.

Imputácia chýbajúcich hodnôt

Na účely riešenia chýbajúcich hodnôt sme testovali viaceré stratégie imputácie (Li et al., 2023), pričom sme kládli dôraz na výber takej metódy, ktorá minimalizuje riziko narušenia pôvodnej štruktúry a logiky segmentácie dátového súboru:

- Priemerná imputácia bola vylúčená z dôvodu jej tendencie skresľovať rozptyl a potenciálne skreslenie modelov založených na segmentoch;
- Multiple Imputation by Chained Equations (MICE) vykazovala mierny výkon pre niektoré premenné a bola zvažovaná v predbežných skúškach;
- Náhodná imputácia lesa bola nakoniec vybraná ako najefektívnejšia metóda, ktorá vyvažuje presnosť so zachovaním premenných vzťahov.

Algoritmus Random Forest bol aplikovaný špeciálne na doplnenie chýbajúcich hodnôt pre technické vlastnosti vozidla, menovite výkon motora, hmotnosť a rok výroby, učením sa vzorov z podobných úplných záznamov.

3.3 Štatistické metódy a modely strojového učenia

Táto fáza analýzy bola zameraná na získanie poznatkov z vyčisteného súboru údajov prostredníctvom štatistického modelovania a techník strojového učenia. Použil sa dvojcestný prístup: štatistická segmentácia založená na kontraste, po ktorej nasledovalo prediktívne modelovanie.

Segmentácia poistencov pomocou kontrastnej analýzy

Na segmentovanie poistencov do štatisticky významných skupín sme použili zovšeobecnené lineárne modely (GLM) v spojení s kontrastnou analýzou. Tento proces sa opieral o výpočet priemerov najmenších štvorcov (LS-means) na neutralizáciu vplyvu mätúcich premenných a na izoláciu účinkov kľúčových segmentačných faktorov.

3.4 Postup segmentácie

Postup segmentácie poistencov prebiehal v niekoľkých analytických krokoch, ktorých cieľom bolo vytvoriť štatisticky homogénne skupiny na základe relevantných rizikových faktorov.

- Identifikácia faktorov segmentácie – vrátane veku vodiča, značky vozidla, geografickej polohy a výkonu motora;
- Aplikácia GLM – na modelovanie závažnosti reklamácie (CS) ako funkcie týchto prediktorov;
- Kontrastná analýza – na testovanie štatisticky významných rozdielov medzi identifikovanými segmentmi;
- Spresnenie segmentov – zlúčenie alebo rozdelenie skupín na základe štatistickej významnosti, aby sa zabezpečila homogenita v rámci segmentov a heterogenita medzi nimi.

Konečná segmentácia priniesla znížený počet štatisticky odlišných zhlukov poistencov a profilov vozidiel, z ktorých každý sa vyznačuje významnými rozdielmi v priemernej závažnosti poistných udalostí.

Modelovanie claim severity pomocou Random Forest

Po segmentácii sme využili Random Forest (Breiman, 2001) model na predikciu claim severity. Tento model bol trénovaný na predikciu výšky poistných škôd na základe:

- veku vodiča,
- výkonu a hmotnosti vozidla,
- značky vozidla,
- histórie škodovosti.

Parametre modelu**:**

- počet stromov: 500,
- hĺbka stromov: Neobmedzená,
- bootstrap vzorkovanie: Áno,
- kritérium rozdelenia: MSE (Mean Squared Error).

Model bol trénovaný na 80 % údajov daného súboru a testovaný na zostávajúcich 20 %, čím sa zabezpečilo, že metriky výkonu odzrkadľujú zovšeobecnenie na neviditeľné údaje.

Výhody použitia Random Forest:

- Dokáže pracovať s veľkými datasetmi a heterogénnymi dátami,
- Minimalizuje multikolinearitu medzi premennými,

- Má vysokú predikčnú presnosť a je odolný voči pretrénovaniu.

Validácia modelov

Na vyhodnotenie výkonnosti štatistických modelov a modelov strojového učenia sa použili viaceré overovacie metriky:

- Pre štatistický model (GLM + kontrastná analýza):
  - P-hodnota faktorov (≤0,05),
  - R^2 a AIC kritérium.
- Pre Random Forest model:
  - Mean Absolute Error (MAE),
  - Root Mean Squared Error (RMSE),
  - Feature Importance Analysis.

Výsledky oboch modelovacích ciest boli porovnané s cieľom identifikovať najefektívnejšiu metodiku pre segmentáciu poistencov a predikciu závažnosti poistných udalostí v kontexte PZP.

4 Výsledky a diskusia

Táto časť prezentuje výsledky segmentácie poistencov, predikciu závažnosti poistnej udalosti a porovnávacie vyhodnotenie štatistických modelov a modelov strojového učenia. Diskusia poukazuje na silné stránky a obmedzenia každého prístupu a identifikuje najvplyvnejšie faktory ovplyvňujúce závažnosť poistných udalostí v kontexte poistenia zodpovednosti za škodu spôsobenú prevádzkou motorového vozidla (PZP).

4.1 Výsledky segmentácie poistencov pomocou kontrastnej analýzy

Segmentácia poistencov bola vykonaná pomocou všeobecných lineárnych modelov GLM v kombinácii s kontrastnou analýzou. Tento prístup umožnil vytvorenie štatisticky homogénnych podskupín na základe vybraných rizikových faktorov.

Hlavné zistenia sú nasledovné:

- *Vek vodiča:*
  - Najvyššia priemerná závažnosť poistnej udalosti bola pozorovaná u vodičov vo veku 25 rokov alebo mladších. Rozdiel v závažnosti reklamácie v porovnaní s vodičmi vo veku 25–55 rokov bol štatisticky významný ($p < 0,01$).
  - Vodiči nad 55 rokov vykazovali mierne nižšiu priemernú závažnosť nárokov ako vodiči v strednom veku, ale tento rozdiel nebol štatisticky významný ($p = 0,12$).
  - Na základe týchto výsledkov boli vodiči zoskupení do troch segmentov:
    - Skupina A (≤ 25 rokov) – Vysoké riziko; zvýšené poistné
    - Skupina B (26–55 rokov) – Stredné riziko
    - Skupina C (56+ rokov) – Nízke riziko
- **Značka vozidla**:
  - Vodiči prémiových/luxusných vozidiel značiek ako BMW, Mercedes a Audi vykazovali vyššiu závažnosť nárokov v porovnaní s vodičmi štandardných značiek ako Škoda, Ford alebo Opel.
  - Rozdiel medzi prémiovými a neprémiovými značkami bol štatisticky významný ($p < 0,05$).

- **Lokalita**:
  - Vodiči z veľkých miest (Bratislava, Košice) mali o 25 % vyššiu závažnosť poistných udalostí oproti vodičom z menších miest a vidieka.
  - Rozdiel bol štatisticky významný **(**$p < 0,01$**).**
- **Výkon a hmotnosť vozidla**:
  - Vyšší výkon motora bol silne pozitívne korelovaný so závažnosťou poistných udalostí ($r = 0,68$**).**
  - Hmotnosť mala mierny negatívny vplyv **(**$r = 0,21$), čo naznačuje, že ťažšie vozidlá spôsobujú závažnejšie škody.

Celkovo sa segmentačný prístup založený na kontrastnej analýze ukázal ako účinnejší pri rozlišovaní rizikových skupín ako tradičné metódy, ktoré sa často spoliehajú len na základné kategorizácie veku alebo výkonu vozidla.

4.2 Výsledky predikcie závažnosti poistných udalostí pomocou Random Forest modelu

Na predikciu výšky škôd bol aplikovaný Random Forest model, ktorý bol trénovaný na 80 % datasetu a testovaný na zvyšných 20 %.

Výsledky modelu:

- MAE (Mean Absolute Error): 312,54 €
- RMSE (Root Mean Squared Error): 528,87 €
- R^2 (koeficient determinácie): 0,76

Tab. 1: Dôležitosť faktorov v modeli Random Forest

Faktor	Relatívna dôležitosť (%)
Výkon motora (kW)	29,3 %
Vek vodiča	23,7 %
Značka vozidla	18,2 %
Lokalita vodiča	14,5 %
Hmotnosť vozidla	7,9 %
História škodovosti	6,4 %

Zdroj: vlastné spracovanie

Model potvrdil, že výkon motora a vek vodiča sú najvýznamnejšie faktory ovplyvňujúce výšku škôd**.** Zároveň ukázal mierne lepšiu predikčnú presnosť v porovnaní s GLM modelom **(**$R^2 = 0,76$ vs. $R^2 = 0,71$ pri GLM**).**

4.3 Porovnanie štatistických a ML prístupov

Hlavné zistenia z porovnania štatistických prístupov a strojového učenia sú zobrazené v tabuľke 2.

Tab. 2: Porovnanie GLM a ML

Kritérium	**GLM + kontrastná analýza**	**Random Forest**
Segmentácia poistencov	Áno, manuálne definované skupiny	Automatická segmentácia
Interpretovateľnosť	Vysoká (jasné vzťahy medzi premennými)	Nízka (black-box model)

Presnosť predikcie (R^2)	0,71	0,76
Práca s veľkými dátami	Mierne obmedzená	Výborná
Schopnosť zachytiť nelineárne vzťahy	Obmedzená	Vysoká

Zdroj: vlastné spracovani

Môžeme ich zhrnúť do nasledujúcich bodov:

- GLM v kombinácii s analýzou kontrastu poskytuje vysokú mieru interpretovateľnosti, čo je kľúčové najmä v regulovaných prostrediach, kde sa vyžaduje transparentnosť a možnosť spätnej validácie modelu;
- Model Random Forest poskytuje vyššiu presnosť predikcie, zachytáva zložité nelineárne vzory a je vhodnejší pre prostredia s veľkým objemom údajov;
- Hybridný prístup – integrácia štatistického modelovania so strojovým učením – prináša najefektívnejšie riešenie, ktoré vyvažuje transparentnosť s presnosťou.

Štatistické modely zabezpečujú teoretickú konzistentnosť a súlad s predpismi, zatiaľ čo modely ML prispievajú k zlepšenej predikcii rizík a prevádzkovej efektívnosti

4.4 Diskusia a implikácie pre prax

Výsledky tejto štúdie naznačujú, že poisťovne môžu výrazne zlepšiť segmentáciu poistencov aj predikciu rizík kombináciou štatistických metodík a metodík strojového učenia. Z tohto výskumu vyplýva niekoľko praktických dôsledkov:

- Optimalizácia cien: Modely vylepšené ML umožňujú poisťovniam nastaviť presnejšie a individualizované sadzby poistného na základe jemne vyladených parametrov rizika.
- Dynamické modelovanie rizík: Využitie strojového učenia umožňuje priebežnú aktualizáciu rizikových profilov na základe aktuálnych údajov, čím sa zvyšuje flexibilita a operatívnosť procesov poistného upisovania.
- Regulačné výzvy: Používanie zložitých modelov ML vyvoláva problémy s vysvetliteľnosťou modelov, ktoré môžu predstavovať problémy v rámci regulačných rámcov, ako je Solventnosť II ((Trombetti, 2017). Kľúčovým hľadiskom zostáva zabezpečenie zrozumiteľnosti a opodstatnenosti výstupov modelu.

Zistenia podporujú zmiešaný analytický prístup, v ktorom sa modely ML primárne používajú na spracovanie údajov, predikčnú analytiku a detekciu vzorov, zatiaľ čo štatistické techniky sú zachované na validáciu, interpretáciu a regulačné vykazovanie. Táto synergia nielen zvyšuje výkonnosť modelu, ale pomáha aj udržiavať dôveru regulačných orgánov aj poistencov.

5 Záver

Táto štúdia sa zamerala na skúmanie integrácie štatistických metód a modelov strojového učenia (ML) na analýzu údajov v oblasti poistenia zodpovednosti za škodu spôsobenú prevádzkou motorového vozidla (PZP). Cieľom bolo posúdiť, ako môže synergia týchto prístupov zlepšiť segmentáciu poistencov, zlepšiť predikciu rizika a podporiť efektívnejšie stratégie oceňovania poistného. Kľúčové zistenia štúdie sú zhrnuté nižšie.

Po prvé, použitie kontrastnej analýzy v kombinácii so všeobecnými lineárnymi modelmi (GLM) umožnilo presnejšiu segmentáciu poistencov do homogénnych rizikových skupín. Začlenením premenných, ako je vek vodiča, značka vozidla a geografická poloha, sa poskytol

hlbší pohľad na rizikové profily a podporil sa vývoj spravodlivejších modelov oceňovania založených na dôkazoch.

Po druhé, algoritmus Random Forest preukázal vyššiu presnosť predikcie ($R^2 = 0,76$) pri modelovaní závažnosti nároku, čím prekonal tradičné štatistické prístupy ($R^2 = 0,71$). Model efektívne zachytil nelineárne vzťahy a identifikoval kľúčové rizikové faktory, najmä výkon motora a vek vodiča, ktoré významne ovplyvňujú výsledky poistných udalostí.

Porovnanie prístupov ukázalo, že kombinácia štatistických modelov a modelov strojového učenia prináša optimálne výsledky. Zatiaľ čo štatistické modely poskytujú interpretovateľnosť, teoretickú konzistentnosť a transparentnosť, ktoré sú nevyhnutné pre súlad s predpismi, techniky strojového učenia prinášajú vynikajúcu predikčnú silu, škálovateľnosť a schopnosť spracovávať veľké a zložité súbory údajov. Tento hybridný rámec podporuje informovanejšie rozhodnutia pri upisovaní poistného a vyššiu efektivitu pri spracovaní údajov. Z praktického hľadiska zistenia naznačujú, že poisťovne môžu výrazne zlepšiť riadenie rizík a cenové procesy prijatím kombinovaného prístupu. Integrácia štatistických a ML metodológií umožňuje presnejšiu segmentáciu, dynamické modelovanie rizík a škálovateľnú analýzu údajov, čo v konečnom dôsledku vedie k presnejším a spravodlivejším prémiovým štruktúram. Prijatie modelov ML však prináša určité regulačné a etické výzvy. Najmä povaha „čiernej skrinky“ mnohých algoritmov ML môže brániť transparentnosti a sťažiť regulačné schvaľovanie, najmä v rámcoch, ako je Solventnosť II alebo IFRS 17. Preto je nevyhnutné zabezpečiť preukázateľnosť a spravodlivosť vo výstupoch modelu.

Pri pohľade do budúcnosti možno identifikovať niekoľko smerov budúceho výskumu. Rozšírenie o ďalšie techniky ML: Ďalšie štúdie by mali preskúmať použitie alternatívnych algoritmov, ako sú neurónové siete alebo stroje na zvýšenie gradientu, na vyhodnotenie ich výkonnosti pri predikcii rizika PZP. Analýza údajov v reálnom čase: Implementácia modelov ML schopných spracovávať streamingové údaje by mohla umožniť poisťovniam dynamicky reagovať na zmeny v individuálnych rizikových profiloch. Etické a regulačné hľadiská: Na zosúladenie aplikácií ML s vyvíjajúcimi sa regulačnými požiadavkami a spoločenskými očakávaniami je nevyhnutné neustále skúmanie ochrany údajov, algoritmickej spravodlivosti a preukázateľnosti modelov.

Na záver, kombinácia štatistickej analýzy a strojového učenia predstavuje účinný prístup k modelovaniu rizík a optimalizácii produktov v poisťovníctve. Keďže priemysel čelí rastúcemu tlaku na presnosť, efektívnosť a dodržiavanie predpisov, táto integrovaná stratégia ponúka cestu ku konkurenčnej výhode a lepšiemu rozhodovaniu. Výsledky tejto štúdie potvrdzujú, že spojenie kontrastnej analýzy s modelmi ML umožňuje robustnú segmentáciu poistencov a presnú predikciu závažnosti poistných udalostí v oblasti PZP. Budúci výskum by mal uprednostňovať pokročilé prediktívne modelovanie, integráciu telematických údajov a vývoj transparentných rámcov strojového učenia kompatibilných s reguláciou.

Príspevok bol spracovaný v rámci riešenia grantovej úlohy VEGA 1/0096/23 Vybrané metódy riadenia rizík pri implementácii čiastkových interných modelov na stanovenie kapitálovej požiadavky na solventnosť.

Literatúra

1. Baudry, M., & Robert, M. (2019). A machine learning approach for individual claims reserving in insurance. *ASMBI, 35*(5), 1127–1155.
2. Breiman, L. (2001). *Machine Learning, 45*(1), 5–32. doi:10.1023/a:1010933404324

3. Chollet, F. (2019). *Deep Learning v jazyku Python: Knihovny Keras, TensorFlow*. TensorFlow. Grada Publishing.
4. Grize, Y.-L., Fischer, W., & Lützelschwab, C. (2020). Machine learning applications in nonlife insurance. *Applied Stochastic Models in Business and Industry*, *36*(4), 523–537. doi:10.1002/asmb.2543
5. James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. Springer.
6. Li, A., Perry, R., Huynh, C., Tomita, T. M., Mehta, R., Arroyo, J., Vogelstein, J. (2023). Manifold oblique random forests: Towards closing the gap on convolutional deep networks. *SIAM Journal on Mathematics of Data Science*, *5*(1), 77–96. doi:10.1137/21m1449117
7. Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. *International Journal of Information Management Data Insights*, *1*(2).
8. Reiff, M., Šoltés, E., Komara, S., Šoltésová, T., & Zelinová, S. (2022). Segmentation and estimation of claim severity in motor third-party liability insurance through contrast analysis. *Equilibrium*, *17*(3), 803–842. doi:10.24136/eq.2022.028
9. Šoltés, E., Zelinová, S., & Bilíková, M. (2019). General linear model: An effective tool for analysis of claim severity in motor third party liability insurance. *Statistics in Transition New Series*, *20*(4), 13–31. doi:10.21307/stattrans-2019-032
10. Trombetti, G. (2017). *Solvency II*. Eurilink.