Predviđanje podnošenja zahteva za naplatu štete od strane nosioca polise osiguranja

Nemanja Todorović, IN 27/2020, nemanjatodorovic132002002@gmail.com

Vladimir Blanuša, IN 28/2020, vlada2002blanusa@gmail.com

I. UVOD

Osiguranje motornih vozila je ključni segment osiguravajuće industrije koji pruža zaštitu od finansijskih gubitaka nastalih usled saobraćajnih nesreća, krađe ili drugih nepredviđenih događaja.

Upravljanje zahtevima za isplatu predstavlja ključni deo procesa osiguranja vozila. Predviđanje podnošenja zahteva za naplatu štete od strane nosioca polise postaje sve važnije s obzirom na dinamičnost tržišta osiguranja i sve veći broj motornih vozila u svetu.

Ovaj istraživački rad fokusira se na primenu tehnika mašinskog učenja kako bi se analiziralo i predvidelo podnošenje zahteva za naplatu štete od strane nosioca polise osiguranja vozila.

U okviru ovog istraživanja primenili smo tri različita različita klasifikatora, a to su kNN (*k-Nearest Neighbors*), stabla odluke (*Decision Tree*) i Naivni Bajesov klasifikator (*Naïve Bayes Classifier*). Svaki od ovih klasifikatora treniran je kako u originalnom prostoru osobina, tako i u prostoru smanjene dimenzionalnosti koji je dobijen primenom PCA *(Principal Component Analysis*) algoritma.

II. BAZA PODATAKA

Baza podataka koju obrađujemo sastoji se od 44 obeležja i 58593 uzorka. Od ukupnog broja obeležja, 15 su numerička, a preostalih 29 su kategorička obeležja.

Jedan uzorak u bazi predstavlja jednu izdatu polisu osiguravajuće kompanije, odnosno podatke o motornom vozilu koje je predmet osiguranja, kao i podatke o vlasniku vozila, gustini naseljenosti grada u kom vlasnik živi… Svaki uzorak sa sobom nosi i informaciju da li je u toku godinu dana vlasnik polise podneo zahtev za isplatu (*is_claim*).

Takođe, vrednosti određenih obeležja su unapred normalizovane i to će nam dosta pomoći u daljem radu.

III. ANALIZA PODATAKA

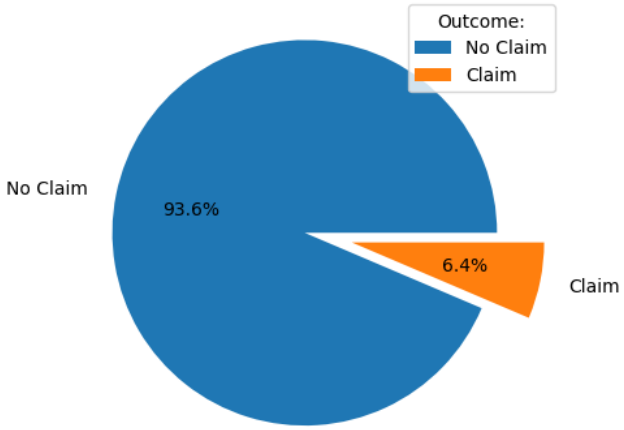
Prilikom analize baze podataka, sproveli smo temeljnu reviziju svih obeležja kako bismo identifikovali njihovu relevantnost za problem koji rešavamo. Ovaj proces uključivao je ispitivanje svih obeležja i uzoraka u bazi podataka radi identifikacije eventualnih nedostataka, višestrukih redudantnosti ili nepravilnosti.

Za početak izbacujemo identifikaciono obeležje uzorka (*policy_id*), budući da nije relevantno za problem koji rešavamo. Pored identifikacionog obeležja uzorka, izbacujemo i obeležja proizvođača automobila (*make*) I model automobila (*model*). Razlog za izbacivanje ova dva obeležja leži u činjenici da su ova obeležja bila predstavljena putem kodnih imena za koja nismo bili u mogućnosti da utvrdimo tačno značenje. Takođe, način na koji će određena marka vozila I sam proizvođač biti uneti u bazu varira od jedne do druge osiguravajuće kuće. Samim tim, ova obeležja ne igraju preveliku ulogu u daljem radu. Poslednje obeležje koje smo izbacili je oznaka za klaster područja (*area_cluster*), a razlog izbacivanja je isti kao i u slučaju prethodna dva izbačena obeležja. Pored toga, još jedan od razloga za izbacivanje ovog obeležja je postojanje obeležja za gustinu naseljenosti (*population_density*) koje dosta preciznije uvodi uticaj mesta stanovanja vlasnika polise u naš problem.

Pri analizi nepravilnih vrednosti u određenim obeležjima, posebnu pažnju posvetili smo obeležjima snage motora (*max_power*) i obrtnog momenta (*max_torque*). Ova obeležja su, zajedno sa obeležjima za dimenzije motornih vozila, od ključnog značaja prilikom analize performansi vozila i potencijalnih rizika. Međutim, prilikom analize podataka, uočili smo da su ova obeležja zabeležena na način koji uključuje i mernu jedinicu zajedno sa samom numeričkom vrednošću. U procesu rešavanja ovog problema, izdvojili smo samo numeričke vrednosti iz celokupnog zapisa vrednosti ovih obeležja, eliminišući pritom mernu jedinicu, a zatim smo celo obeležje pretvorili u numerički tip obeležja. Ovako izmenjeno obeležje će dalje imati veliki uticaj na kvalitet rada samih klasifikatora i na kvalitet rezultata klasifikacije.

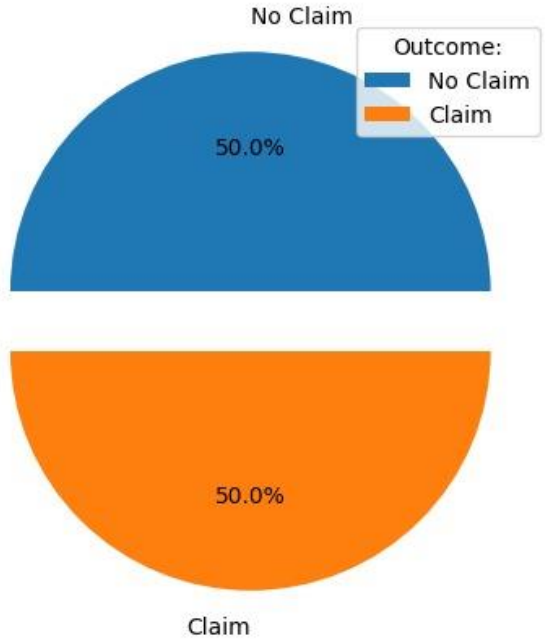
Nakon prvog dela analize podataka, broj uzoraka je ostao isti - 58593, dok je broj obeležja spao na 41.

U drugom delu analize podataka bavili smo se zastupljenošću uzoraka u svakoj od klasa obeležja koje nam govori da li je vlasnik polise podneo zahtev za isplatu (*is_claim*), koje je istovremeno i obeležje čiju vrednost predviđamo. Odnos zastupljenosti klasa je prikazan na slici 1.1, gde se jasno vidi da postoji ogromna razlika u broju uzoraka koji pripadaju različitim klasama. Sa posledicama nebalansiranosti klasa smo se upoznali pri samom početku kreiranja klasifikacionih modela, kada smo naišli na izuzezno niske rezultate.



Slika 1.1 – Odnos zastupljenosti klasa pre obrade

Kako bismo izbegli različite probleme koje ovakva raspodela može da nam donese u daljem obučavanju modela, koristili smo SMOTE (*Synthetic Minority Over-Sampling Technique*) tehniku. Ovom tehnikom generisali smo uzorke manjinske klase kombinovanjem karakteristika sličnih uzoraka te klase i time smo postigli balansiranost između klasa. Zastupljenost klasa nakon primene ove metode data je na slici 1.2



Slika 1.2 – Odnos zastupljenosti klasa nakon obrade

Ovako uređene podatke dalje delimo u skupove za trening i test kojima ćemo kasnije raditi obučavanje modela i predikciju. U test skup smo smestili 20% uzoraka kako bi obezbedili adekvatan broj uzoraka za evaluaciju performansi modela.

IV. UNAKRSNA VALIDACIJA

Pre samog obučavanja modela klasifikacije, potrebno je izvršiti odabir optimalnih hiperparametara koji će dovesti do toga da naši modeli daju najbolje moguće rezultate.

U cilju pronalaženja optimalnih hiperparametara, sproveli smo unakrsnu validaciju pre obučavanja svakog od modela. Ovaj proces uključuje iterativno procenjivanje performansi modela koristeći različite kombinacije hiperparametara, pri čemu smo uzeli u obzir različite metrike evaluacije, kao i varijacije u broju suseda. Kao evaluacionu metriku tokom pretrage hiperparametara koristili smo *f1 score,* jer smo utrvrdili da je ona najpreporučljivija metrika za problem binarne klasifikacije i na osnovu nje smo vršili odabir optimalnih hiperparametara.

Unakrsna validacija odvijala se u pet iteracija kako bi se osigurala pouzdanost i preciznost rezultata. Nakon završene validacije, analizirali smo dobijene rezultate i za svaki klasifikator smo odabrali onu kombinaciju hiperparametara koja je dala najbolje rezultate.

V. kNN

Za prvi model klasifikacije koristili smo kNN algoritam (*k-Nearest Neighbors*) na tri različita načina. Prvo smo primenili standardni kNN, zatim kNN sa uvođenjem standardizacije obeležja, a potom i kNN sa uvođenjem normalizacije obeležja. Korišćena je *Manhattan* metrika, a broj suseda je postavljen na 1 u svakom od modela, budući da su ovi hiperparametri dali najbolje rezultate tokom unakrsne validacije.

Nakon primene ovih modela, dodatno smo istražili i kNN u prostoru smanjene dimenzionalnosti, koristeći PCA *(Principal Component Analysis*) algoritam. Optimalni hiperparametri za ovaj model bili su 20 komponenti i 1 sused, takođe dobijeni unaksrnom validacijom. Kao meru uspešnosti ponovo smo posmatrali *f1 score*.

U poređenju rezultata predikcije za svaki od kNN modela, primećeno je da se prvi klasifikator, bez primene standardizacije I normalizacije, u originalnom prostoru obeležja, pokazao kao najbolji u smislu prediktivne tačnosti. Samim tim, u daljem istraživanju ćemo porediti rezultate svih ostalih modela upravo sa ovim kNN modelom.

VI. STABLA ODLUKE

U ovom poglavlju, fokusiramo se na obučavanje modela za klasifikaciju korišćenjem algoritma stabla pretrage (*Decision Tree*).

Kao i u prethodnom slučaju sa kNN modelima, prvi korak je sprovođenje unakrsne validacije radi identifikacije optimalnih hiperparametara. Za razliku od prethodnih modela, u ovom delu obučićemo dva modela, jedan u originalnom prostoru obeležja i drugi u prostoru smanjene dimenzionalnosti. Ponovo radimo unakrsnu validaciju u pet iteracija i za optimalne hiperparametre

dobijamo *Entropy* kriterijum i 1024 za nasumičnost estimatora (*random_state*).

Nakon što smo dobili optimalne hiperparametre izvršili smo obuku modela sa dobijenim hiperparametrima, a potom i na predikciju.

Ceo ovaj proces ponovili smo i u prostoru smanjene dimenzionalnosti gde smo, nakon unakrsne validacije, dobili skoro identične hiperparametre.

Nakon predikcije oba modela, izvršili smo detaljno poređenje dobijenih rezultata. Na osnovu dobijenih rezultata oba modela zaključujemo da je model obučen u originalnom prostoru obeležja dao u maloj meri bolje rezultate, pa ćemo taj model i koristiti u krajnjem poređenju najefikasnijih modela.

VII. NAIVNI BAJESOV KLASIFIKATOR

U ovom poglavlju bavili smo se Naivnim Bajesovim klasifikatorom (*Naïve Bayes Classifier*), koji predstavlja poslednji model koji smo analizirali u okviru ovog istraživanja. Kao i kod prethodna dva klasifikatora, prvo smo izvršili unakrsnu validaciju, pri čemu smo ovoga puta koristili 3 iteracije umesto 5 kao kod prethodnih modela. Time smo uštedeli na vremenu izvršavanja unakrsne validacije, jer sam način rada Naivnog Bajesov klasifikatora, zasnovan na Gausovoj raspodeli, zahteva veliki broj tačaka raspoređenih u prostoru čija obrada zahteva dosta vremena.

Nakon dobijanja optimalnih hiperparametara, prešli smo na treniranje modela i izvršili predikciju. Oba postupka ponovili smo i u originalnom prostoru i u prostoru smanjene dimenzionalnosti. Poređenjem rezultata dobijenih iz oba prostora, mogli smo da procenimo performanse modela i identifikujemo najefikasniji, a to je, u slučaju ovog klasifikatora, bio model obučen u originalnom prostoru obeležja.

VIII. POREĐENJE NAJBOLJIH MODELA

U prethodnom segmentu istraživanja, obučili smo tri modela za klasifikaciju: kNN u originalnom prostoru obeležja i u prostoru smanjene dimenzionalnosti, sa i bez standardizacije i normalizacije; Stablo odluke u orignialnom prostoru i u smanjenom prostoru; i Naivni Bajesov klasifikator u originalnom i smanjenom prostoru obeležja. Prilikom obuke svakog od ova tri modela vršili smo međusobna poređenja i identifikovali smo tri najefikasnija za svaki klasifikator: kNN u originalnom prostoru obeležja, bez standardizacije i normalizacije; Stablo odluke u originalnom prostoru obeležja; i Naivni Bajesov klasifikator u originalnom prostoru obeležja.

Sada je neophodno izvršiti detaljno poređenje ova tri modela kako bismo doneli konačan sud o tome koji je najbolji. Poređenjem *f1 score* predikcije za sve tri klasifikacije, došli smo do pobedničkog modela.

IX. POBEDNIČKI MODEL

Nama, zbog svoje jednostavnosti, neočekivani pobednik, Stablo odluke, svoje prvo mesto među našim klasifikatorima obezbedio je visokim rezultatima rasprostranjenim sa svim poljima . Razlog za to smatramo da je postojanje šablona u našem *dataset*-u koje Stablo odluke najbolje može da iskoristi.

	precision	recall	f1-score	support
0	0.89	0.88	0.88	10969
1	0.88	0.90	0.89	10969
accuracy			0.89	21938
macro avg	0.89	0.89	0.89	21938
weighted avg	0.89	0.89	0.89	21938

Slika 2 – Rezultati klasifikatora Stabla odluka

Preciznost (*precision*) od 0.89 za klasu 0 i 0.88 za klasu 1 pokazuje da model ima visoku tačnost u predviđanju prave klase.

Pokrivenost (*recall*) od 0.88 za klasu 0 i 0.90 za klasu 1 pokazuje da model dobro prepoznaje prave pozitivne instance obe klase.

F1 score, koji je harmonička sredina između preciznosti i pokrivenosti, takođe je visok za obe klase, što znači da model dobro balansira između tačnosti i pokrivenosti, ne ostvarujući jedno na račun drugog.

X. ZAKLJUČAK

Naše istraživanje povtvrđuje da postoji potreba za prilagođavanjem tradicionalnih metoda upravljanja zahtevima za isplatu štete uz pomoć naprednih tehnika mašinskog učenja. Kroz primenu metoda kao što su kNN klasfikator, Stabla odluke, Naivni Bajesov klasfikator, SMOTE tehnika, uspeli smo identifikovati modele koji su efikasni u predviđanju podnošenja zahteva za isplatu, čime pružamo osiguravajućim kompanijama važan alat za optimizaciju procesa osiguranja motornih vozila.

U skladu sa tim, ovaj rad predstavlja značajan korak napred u razumevanju i unapređenju procesa osiguranja motornih vozila kroz primenu naprednih tehnika mašinskog učenja. Budući rad može biti usmeren na proširenje analize na druge vrste osiguranja ili na dalje unapređenje modela kroz korišćenje naprednih tehnika obrade podataka za ostvarivanje konkurentske prednosti u ovakvoj industriji.