Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/15BE40Skipina

PREDIKCIJA RIZIKA I KLASIFIKACIJA OZBILJNOSTI SUDARA MOTORNIH VOZILA U NJUJORKU

TRAFFIC ACCIDENT RISK PREDICTION AND SEVERITY CLASSIFICATION IN THE CITY OF NEW YORK

Milica Škipina, *Fakultet tehničkih nauka, Novi Sad*

Oblast – RAČUNARSTVO I AUTOMATIKA

Kratak sadržaj – *Svake godine u saobraćajnim nesrećama na putevima širom svijeta pogine 1,35 i bude povrijeđeno 20-50 miliona ljudi, što znači da svakog dana u prosjeku skoro 3.700 ljudi izgubi život u saobraćaju. Više od polovine poginulih su pješaci, motoristi ili biciklisti. U ovom radu će biti analizirani faktori koji direktno utiču na povećanje vjerovatnoće pojave sudara motornih vozila, a zatim predstavljena metodologija za predviđanje rizika kao i klasifikaciju nivoa ozbiljnosti sudara korištenjem klasifikacionih modela mašinskog učenja. Predloženi model objedinjuje podatke o sudarima, ulicama Njujorka, protoku saobraćaja na pojedinim dionicama i podatke o vremenu i može se koristiti za identifikaciju gdje i kada je rizik od nesreće značajno veći od prosjeka kako bi se preduzele radnje za smanjenje tog rizika. Rezultati modela za predikciju sudara dostižu tačnost od 70%, dok model za klasifikaciju ozbiljnosti sudara postiže makro-prosječni F1-skor od 0,56.*

Ključne reči: *saobraćajne nesreće, mašinsko učenje, klasifikacioni algoritmi, urbanističko planiranje*

Abstract – *Annually 1.35 million people worldwide die in road traffic, while 20-50 million get injured. That means that every day, almost 3,700 people are killed globally in crashes. More than half of those killed are pedestrians, motorcyclists, or cyclists. In this paper, we analyze the factors that directly affect the risk of motor vehicle crashes and propose the methodology for predicting traffic accidents and classification of collision severity. We were using various machine learning classification algorithms. We present a dataset obtained by extracting accident, road, traffic, and weather-related information from various data sources. The proposed model can identify the time and place when the risk of traffic accidents is higher so that risk can be reduced with proper actions. Experimental results show that the traffic accident prediction model can reach an accuracy of 70%, while the collision severity classification model achieves a macro-average F1-score of 0.56.*

Keywords: *traffic accidents, machine learning, classification algorithms, urban planning*

1. UVOD

Danas ljudi širom svijeta u cilju obavljanja svakodnevnih aktivnosti poput odlaska na posao ili u kupovinu za prelazak sa jedne na drugu lokaciju koriste različita prevozna sredstva. Posljedica toga je veliki broj sudara. Svake godine u saobraćajnim nesrećama u kojima učestvuju automobili, autobusi, motocikli, bicikli, kamioni ili pješaci na putevima pogine 1,35 i bude povrijeđeno 20-50 miliona ljudi širom svijeta. Više od polovine poginulih su pješaci, motoristi ili biciklisti. Na svaku osobu koja smrtno strada od posljedica povreda izazvanih u saobraćajnom udesu, na desetine ljudi koji su preživjeli za posljedicu imaju kratkoročni ili trajni invaliditet koji može rezultovati stalnim ograničenjima fizičkog funkcionisanja, psihosocijalnim posljedicama ili smanjenim kvalitetom života [1]. U Sjedinjenim Američkim državama (SAD), saobraćajne nesreće predstavljaju vodeći uzrok smrti osoba starosti od 19-54 godine i vodeći uzrok neprirodne smrti državljana SAD koji borave ili putuju u inostranstvo [2].

Saobraćajne nesreće ostavljaju i ekonomske posljedice. Procjenjuje se da će povrede sa ili bez tragičnog ishoda nastale u saobraćajnim udesima u periodu od 2015. do 2030. godine koštati svjetsku ekonomiju otprilike $1,8 triliona američkih dolara [3].

Uzimajući sve ovo u obzir, jako je bitno razmotriti sve moguće načine na koje bismo mogli spriječiti sudare ili reagovati efikasno u slučaju da do njih dođe. Pristup tačnim i ažuriranim informacijama o trenutnoj situaciji na putevima omogućava vozačima, pješacima i putnicima da donose bolje odluke prilikom kretanja u saobraćaju.

U ovom radu će biti analizirani faktori koji direktno utiču na povećanje vjerovatnoće pojave sudara motornih vozila, a zatim predstavljena metodologija za predviđanje kao i klasifikaciju nivoa sudara na sudare sa smrtnim ishodom i /ili ozbiljnim povredama i sudare u kojima je samo došlo do materijalne štete. Rješenje predstavljeno u ovom radu objedinjuje podatke iz različitih izvora: podatke vezane za sudare, razne informacije o ulicama Njujorka, informacije o protoku saobraćaja, kao i informacije o vremenskim prilikama za određene trenutke. Kako se problem predikcije sudara može posmatrati kao problem binarne klasifikacije gdje pozitivna klasa odgovara zabilježenom sudaru na određenoj lokaciji i u određenom trenutku, i za predikciju i za klasifikaciju ozbiljnosti sudara korišteni su različiti klasifikacioni algoritmi: *Random Forest, Extra Trees, LightGBM, XGBoost, CatBoost* i K najbližih komšija (*eng. K nearest neighbors*). Takođe je isproban

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bila dr Jelena Slivka, vanr. prof.

ansambl gdje su kombinovani različiti modeli u cilju poboljšanja performansi.

2. PRETHODNA RJEŠENJA

U [4] su korištena tri javno dostupna skupa podataka obezbijeđena od strane grada Montreala i vlade Kanade koji obuhvataju informacije o nesrećama (datum, vrijeme i lokacija), geometriji puteva i podatke o vremenskim prilikama koji su mjereni na svakih sat vremena na različitim meteorološkim stanicama.

Za kreiranje modela za predikciju sudara korištene su tehnike za analizu i obradu velikih skupova podataka (*eng. Big Data*). Problem neuravnoteženosti skupa podataka je rješen tako što su za negativne primjere izdvojili 0.1% nasumičnih uzoraka od 2,3 milijarde mogućih kombinacija vremenskih trenutaka i segmenata. Obučavali su tri modela bazirana na stablima: *Random Forest, Balanced Radnom Forest i XGBoost*. Kao najbolji se pokazao *Balanced Random Forest* koji tačno detektuje 85% sudara, uz preciznost od 28% i *false positive rate* od 13% na test skupu.

Faktori koji su identifikovani kao najvažniji prilikom predviđanja nesreća su: broj nesreća koje su se prethodnih godina dogodile na određenom segmentu puta, temperatura, dan u godini, sat i vidljivost na putu.

Cilj u [5] je pronalazak nove metode za izbor parametara za problem predikcije saobraćajnih nesreća u realnom vremenu. Poređena su dva načina za izbor parametara: *Random Forest* i *Frequent Pattern Tree* (FP). Podaci koji su korišteni obuhvataju informacije o saobraćajnim nesrećama koje su se dogodile na međudržavnom autoputu I-64 u Virdžiniji 2005. godine, podatke o vremenu, vidljivosti na putu, gustini saobraćaja, brzini kretanja i zauzetosti. Rezultati pokazuju da se FP pokazao kao bolji pri izboru parametara bez obzira na vrstu modela koja je korištena za predviđanje sudara, a u kombinaciji sa modelom *Bayesian network* može se predvidjeti 61,11% nesreća sa *false positive rate* od 38,16%. Najznačaniji atributi su osobine vezane za obim saobraćaja, dok su osobine vezane za brzinu rangirane mnogo niže.

Autori u [6] su razvijali model koji u realnom vremenu predviđa vjerovatnoću da će se dogoditi nesreća sa različitim nivoom ozbiljnosti i došli su do zaključka da različiti faktori utiču na pojavu nesreća različite ozbiljnosti. Korišteni su podaci prikupljeni na 29 milja dugom autoputu I-880 u Kaliforniji i podaci dobijeni sa pet meteoroloških stanica koje se nalaze na udaljenosti od oko 8km od autoputa. Za problem klasifikacije, posmatrana su tri nivoa ozbiljnosti sudara: nesreće sa smrtnim ishodom ili težim povredama, nesreće u kojima je bilo povrijeđenih bez smrtnih slučajeva i nesreće koje za posljedicu imaju samo materijalnu štetu (PDO). Rezultati pokazuju da su se PDO češće dešavale u uslovima kada je bio zagušen saobraćaj, sa promjenljivom brzinom i čestim promjenama trake, dok su se druge dvije grupe nesreća češće dešavale pri manjem zagušenju saobraćaja. Došli su do zaključka da velika brzina zajedno sa velikom razlikom u brzinama između susjednih traka povećava vjerovatnoću pojave fatalnih sudara.

3. SKUPOVI PODATAKA

Kako bi se što preciznije odredila vjerovatnoća pojave nesreće, kao i nivo ozbiljnosti nesreće, korištena su tri javno dostupna skupa podataka obezbijeđena od strane grada Njujorka (*NYC Open Data*) koji sadrže podatke o sudarima, informacije o pojedinačnim segmentima ulica i protoku saobraćaja na njima. Pored toga, pomoću API (*Application Programming Interface*) koj pruža *IBM Weather* [7] preuzete su informacije o vremenu. Svaki od skupova podataka će biti detaljnije opisan u nastavku.

3.1. Sudari motornih vozila

Motor Vehicle Collisions - Crashes [8] sadrži informacije svih policijskih izvještaja o sudarima motornih vozila u Njujorku koji se moraju popuniti u slučaju sudara gdje ima povrijeđenih ili smrtno stradalih osoba ili gdje postoji materijalna šteta u iznosu od najmanje $1.000. Dostupni su podaci od 01.07.2012. godine i ažuriraju se na dnevnom nivou. Svaki red u tabeli se odnosi na jedan sudar i sadrži jedinstven ID sudara, informacije o lokaciji, datumu i vremenu kada se desio sudar, naziv gradske oblasti (*eng. borough*), poštanski broj (*eng. ZIP code*), naziv ulice, najbliže raskrsnice, adresu, broj poginulih i povrijeđenih ljudi, kao i faktore koji su uticali na to da dođe do sudara i tipove vozila. Na slici 1 je prikazana mapa Njujorka sa crvenim tačkama koje označavaju lokacije na kojima je došlo do sudara.

3.2. LION

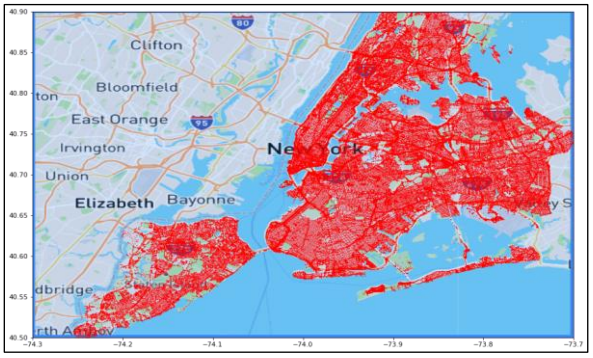
Linearno integrisana uređena mreža (*eng. Linear Integrated Ordered Network – LION*) [9] pomoću jedne linije predstavlja gradske ulice Njujorka i druge linearne karakteristike kao što su obale, željeznice i šetališta zajedno sa nazivima obilježja i opsegom adresa za svaki segment ulice koji se može adresirati. Za svaki segment ulice sadrži naziv ulice u kojoj se nalazi koji je iskorišten za popunjavanje nedostajućih vrijednosti kada je u [8] poznata lokacija, ali nije upisan naziv ulice. Pored naziva ulice, svaki segment je opisan informacijama poput smjera u kojem se odvija saobraćaj, širine, ograničenja brzine u miljama na sat, oznake da li segment pripada mreži ruta za bicikle ili kamione, prioritet za čišćenje snijega, broj saobraćajnih, parking i ukupnih traka na cesti, oznake da li je segment dostupan za pješake ili nije, kao i prostorne koordinate početka i kraja segmenta.

3.3. Protok saobraćaja

Traffic Volume Counts (2014 – 2019) [10] sadrži informacije o broju vozila koja su prošla određenim segmentom puta na svakih sat vremena određenog datuma. Podaci su dostupni samo za 422 datuma u periodu od 13.09.2014. do 24.11.2019. godine.

3.4. IBM Weather API

Informacije o vremenskim prilikama na svakih sat vremena za određene datume su preuzete pomoću *IBM Weather API*-ja. Dobavljene informacije uključuju temperaturu vazduha, subjektivni osjećaj, te količinu rose, vlažnosti, tip i jačinu vjetra kao i udara vjetra, vazdušni pritisak, količinu padavina, uslove (vedro, oblačno, kišovito, oluja,…), oznaku da li je dan ili noć i tip oblaka.



Slika 1. *Mapa Njujorka sa označenim lokacijama na kojima je došlo do sudara*

4. METODOLOGIJA

U ovom poglavlju će biti opisani izazovi sa kojima smo se susreli prilikom rješavanja opisanih problema i način na koji smo ih prevazilazili.

4.1. Integracija i pretprocesiranje podataka

Kako se opsezi datuma dostupnih podataka za [8] i [10] razlikuju, prvo su podaci o sudarima isfiltrirani i zadržani su samo oni koji su se dogodilli u periodu od 13.09.2014. do 24.11.2019. godine. Pošto lokacija predstavlja ključno obilježje za spajanje tri skupa podataka, odbačeni su svi uzorci kojima je ovaj atribut nedostajao. Takođe su izbačeni i redovi koji ni su sadržali informacije o broju poginulih i/ili povrijeđenih osoba jer se ovi atributi koriste prilikom labeliranja podataka za problem klasifikacije ozbiljnosti sudara.

Pošto različiti segmenti unutar jedne ulice mogu imati različite osobine (ograničenje brzine, broj saobraćajnih traka,…), za svaki red iz skupa podataka je na osnovu lokacije pronađen najbliži segment iz ulice u kojoj je došlo do sudara. Zatim su na osnovu njega dobavljene i ostale informacije vezane za taj segment. Spajanje podataka o sudarima sa informacijama o vremenskim prilikama je izvršeno na osnovu sata i datuma u kojem se desio sudar. Rezultujući skup podataka u sebi sadrži ukupno 1.025.439 zapisa o saobraćajnim nesrećama.

Prije spajanja sa skupom podataka o protoku saobraćaja, iz ovog skupa su usrednjene informacije o protoku sobraćaja za svaki od smjerova datog segmenta. Kako ovaj skup podataka sadrži informaije za samo 422 datuma, a postoje i razlike u oznakama segmenata, nakon spajanja sa ostalim skupovima podataka dobili smo novi koji u sebi sadrži 22.872 zapisa o saobraćajnim nesrećama. Kako je u prethodnim radovima zaključeno da ovo predstavlja jedno od najvažnijih obeležja, naše modele smo trenirali na dva odvojena skupa podataka: jedan koji uključivao podatke o protoku saobraćaja i drugi bez njih.

Iz dobijenih skupova podataka su izbačene sve kolone koje su imale veliki broj nedostajućih vrijednosti. Iako datum i vrijeme mogu sadržati informacije korisne za model, one ipak mogu biti neupotrebljive u standardnom formatu („DD-MM-YYYY HH:mm“), pa smo odlučili da u posebne kolone izdvojimo podatke o mjesecu, godini i satu, dok je dan transformisan u kolonu koja predstavlja dan u sedmici. Kako novi atributi prirodno predstavljaju ciklične podatke (podaci koji predstavljaju najudaljenije tačke u jednodimenzionalnoj ravni su u stvari najbliži, npr. 00h i 23h), transformisali smo ih u dvije dimenzije pomoću sinusne i kosinusne transformacije. Takođe je lokacija koja u našem skupu podataka predstavljena pomoću geografske širine i dužine mapirana na *x, y i z* koordinade.

Za kategorička obilježja je korišten *one-hot encoding*, dok su za kolonu koja predstavlja uslove na putu prethodno grupisane vrijednosti u tri grupe: normalni uslovi na puti ili oblačno vrijeme, uslovi sa slabijom vidljivošću i uslovi opasni za vožnju.

4.2. Labeliranje i neuravnoteženost skupa podataka

Za problem klasifikacije ozbiljnosti sudara, svaki od uzoraka je pridružen jednoj od klasa: 1 – ukoliko je u nesreći bilo povrijeđenih ili poginulih osoba i 0 u suprotnom. Kako je broj uzoraka koji pripadaju klasi 1 znatno manji od onih koji pripadaju klasi 0, prije treniranja je izvršen *resampling* podataka, odnosno, prvo je nasumično izbačen određeni broj uzoraka koji pripadaju klasi 0, a zatim je pomoću SMOTE (*Synthetic Minority Oversampling TEchnique*) algoritma povećan broj uzoraka klase 1.

Kod problema predikcije sudara, pozitivnoj klasi odgovaraju svi uzorci iz skupa podataka. Da bi se mogao obučiti model, prethodno je bilo potrebno izgenerisati negativne primjere. Za svaki uzorak iz skupa podataka, generisan je po jedan negativni primjer koji je imao isti datum dok su ulica i vrijeme birani nasumično pod uslovom da se u datoj ulici istog dana nije desio nijedan sudar.

4.3. Obučavanje modela

Prije obučavanja, skup podataka je podijeljen na trening i test skupove u odnosu 90% za trening i 10% za test skup. Pošto oba posmatrana problema predstavljaju problem binarne klasifikacije, obučavani su modeli zasnovani na stablu (*Random Forest, Extra Trees, LightGBM, XGBoost* i *CatBoost*). Zbog načina na koji je izvršen resampling podataka, pored njih je isproban i model K najbližih komšija.

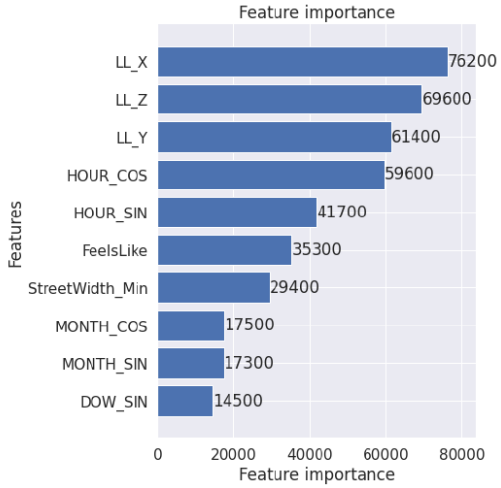
5. Evaluacija rješenja i rezultati

Za evaluaciju modela predikcije korištene su metrike: preciznost (*eng. precision*), tačnost (*eng. accuracy*), *recall* i *F1*-skor, dok su za problem klasifikacije ozbiljnosti sudara korištene makro-prosječne vrijednosti ovih metrika (osim tačnosti). U tabeli 1 su prikazani rezultati makro prosječnog F1-skora svih modela obučenih na skupu podataka bez informacija o protoku saobraćaja.

Tabela 1. *F1-skor modela obučenih na skupu podataka bez informacija o protoku saobraćaja*

Model	Predikcija sudara	Klasifikacija ozbiljnosti
Extra Trees	0,64	0,53
LightGBM	0,68	**0,56**
XGBoost	0,69	**0,56**
CatBoost	**0,70**	0,55
KNN	0,46	0,52
Ensemble	0,69	**0,56**

Kao najbolji model za predikciju sudara se pokazao *CatBoost* koji dostiže F1-skor od 0,70, dok su kod klasifikacije ozbiljnosti sudara sličan F1-skor od 0,56 ostvarila dva modela – *LightGBM* i *XGBoost*. Dodatno je isproban ansambl dva najbolja modela za svaki od problema, ali nije došlo do poboljšanja performansi. Analizom rezultata je utvrđeno da modeli bazirani na stablu postižu približno slične performanse. KNN se pokazao kao znatno lošiji za problem predikcije sudara, dok se manja razlika primjećuje za model koji je istreniran za klasifikaciju ozbiljnosti sudara. Razlog za ovo je vjerovatno SMOTE tehnika za generisanje novih primjera manjinske klase koji u osnovi koristi KNN. Na slici 2 je dat prikaz top 10 najznačajnijih atributa *LightGBM* modela za klasifikaciju ozbiljnosti sudara, gdje se može vidjeti da najvažnije atribute prilikom predikcije predstavljaju lokacija i vrijeme sudara, zatim subjektivni osjećaj, širina ulice, mjesec i dan u sedmici.



Slika 2. *Mapa Njujorka sa označenim lokacijama na kojima je došlo do sudara*

Kako bismo povećali performanse, za problem klasifikacije ozbiljnosti sudara je istreniran model nad skupom podataka koji sadrži informacije o protoku saobraćaja. Kao najbolji model se pokazao *Extra Trees* koji dostiže makro-prosječan F1-skor od 0,69. Iako je ovo značajno poboljšanje, rezultati nisu direktno uporedivi jer ovaj skup podataka sadrži mnogo manje uzoraka u poređenju sa prvim.

5. ZAKLJUČAK

U ovom radu je predstavljeno rješenje za predviđanje lokacije i trenutka u kojem je povećan rizik pojave sudara, kao i klasifikacija ozbiljnosti sudara motornih vozila u Njujorku. Objedinjeni su skupovi iz različitih izvora podataka nad kojim su zatim trenirani različiti klasifikacioni modeli. Nakon rješenja problema neuravnoteženosti skupa podataka, rezultati pokazuju da najbolje performanse prilikom klasifikacije ozbiljnosti sudara postižu modeli *LightGBM* i *XGBoost* sa F1-skorom od 0,56. Za problem predikcije sudara, prvo je bilo neophodno generisati podatke koji se odnose na lokaciju i trenutak u kojem nije došlo do sudara. Nakon treniranja i evaluacije modela, utvrđeno je da najbolje performanse postiže *CatBoost* sa F1-skorom od 0,70. Dodavanje podataka vezanih za protok saobraćaja značajno poboljšava rezultate modela za klasifikaciju nivoa ozbiljnosti sudara, ali rezultati nisu robusni jer su modeli trenirani na znatno manjem skupu podataka.

Predviđanjem težine nesreće moglo bi se efikasnije reagovati u hitnim slučajevima. Na pojavu sudara utiču i drugi faktori kao što je ponašanje vozača, što je teško detektovati kako bi se moglo iskoristiti kao jedan od dodatnih atributa, ali bi modeli mogli biti poboljšani upotrebom podataka vezanih za gustinu naseljenosti određenog mjesta ili pojedinačnih dijelova grada.

6. LITERATURA

[1] Peden, Margaret. (2004). World Report on Road Traffic Injury Prevention.

[2] Centers for Disease Control and Prevention (CDC), National Center for Injury Prevention and Control (NCIPC). Web-based Injury Statistics Query and Reporting System (WISQARS). Available from URL: http://www.cdc.gov/injury/wisqars

[3] Chen, Simiao & Kuhn, Michael & Prettner, Klaus & Bloom, David.(2019). The global macroeconomic burden of road injuries: estimates and projections for 166 countries. The Lancet Planetary Health. 3. e390-e398. 10.1016/S2542-5196(19)30170-6.

[4] A. Hébert, T. Guédon, T. Glatard and B. Jaumard, "High Resolution Road Vehicle Collision Prediction for the City of Montreal," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 1804-1813.

[5] L. Lin, Q. Wang, and A. W. Sadek, "A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction," Transportation Research Part C: Emerging Technologies, vol. 55, pp. 444 – 459, 2015

[6] Xu C, Tarko AP, Wang W, Liu P. Predicting crash likelihood and severity on freeways with real-time loop detector data. Accid Anal Prev. 2013 Aug;57:30-9. doi: 10.1016/j.aap.2013.03.035. Epub 2013 Apr 6. PMID: 23628940.

[7] https://www.ibm.com/weather

[8] https://data.cityofnewyork.us/browse?Data-Collection_Data-Collection=Motor+Vehicle+Collisions

[9] https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-lion.page

[10] https://data.cityofnewyork.us/Transportation/Traffic-Volume-Counts-2014-2019-/ertz-hr4r

Kratka biografija:

Milica Škipina rođena je 1997. godine u Srbinju. Master rad na Fakultetu tehničkih nauka iz oblasti Elektrotehnike i računarstva odbranila je 2021. godine.

kontakt: skipinamilica@gmail.com