# Primena mašinskog učenja za predviđanje rizičnih kupovina automobila na aukcijama

Nina Omerović, Marko Mitrović, Milutin Mirković

Avgust 2024.

## Uvod

U ovom projektnom radu, analizirali smo primenu modela mašinskog učenja na skup podatka koji se odnosi na kupovinu vozila na aukcijama, preuzet sa *Kaggle* takmičenja *Don't Get Kicked!* iz 2012. godine. Cilj projekta bio je razvoj modela koji može detektovati vozila lošijeg kvaliteta nego što se čini na prvi pogled, kako bismo smanjili rizik od kupovine automobila sa skrivenim problemima. U tu svrhu ispitali smo I primenili različite modele I tehnike mašinskog učenja.

Naša analiza je zasnovana na CRISP-DM metodologiji, standardizovanom procesu za izvođenje projekata otkrivanja zakonitosti u podacima, koja uključuje korake poput razumevanja poslovnog problema, razumevanja podataka, pripreme podataka, modeliranja, evaluacije i implementacije, pri čemu su ključni koraci bili razumevanje poslovanja i podataka, kao i priprema podataka, dok smo za kreiranje modela koristili programski jezik *Python* I razvojno okruženje *Jupyter Notebook*.

## Razumevanje poslovnog problema

Karvana (engl. *Carvana*) je onlajn prodavac polovnih automobila. Ima poslovnice širom Sjedinjenih Američkih Država, ali se većina prodaja obavlja putem interneta. Na njihovom sajtu je moguće detaljno razgledati automobile - izgled unutra I spolja, istoriju kvarova i ostale bitne informacije koje korisnicima omogućavaju da donesu informisane odluke bez potrebe da automobil vide uživo.

Karvana vozila nabavlja na aukcijama, jer tu mogu da kupe veliki broj vozila po nižim cenama nego na tržištu. Zatim ih po potrebi sređuju,popravljaju, testiraju I potom prodaju na sajtu po većoj ceni. Problem koji se nameće je nemogućnost uveravanja da je svako vozilo koje se kupuje isplativo, zbog nedostatka informacija I potencijalnih prevara, poput pomeranja kilometraže.

Zbog toga je kompaniji potreban sistem za podršku odlučivanju, koji će, na osnovu dostupnih informacija, detektovati neisplativo vozilo (engl. *Kick*). Ukoliko kupe takvo vozilo, pored promašene investicije, snoše troškove

transporta, popravke I finansijskog gubitka prilikom prodaje. Takođe, reputacija je od suštinskog značaja, jer narušen ugled može da dovede do povećane skeptičnosti I gubitka kupaca.

Prvi poslovni cilj ovog projekta je identifikacija ključnih faktora koji vozilo čine rizičnom investicijom. U tu svrhu, nameravamo da sprovedemo eksplorativnu analizu podataka. Ispitaćemo distribucije podataka, međusobne odnose između atributa, kao i ključne karakteristike koje mogu značajno uticati na odluku o tome da li je vozilo rizična investicija.

Drugi poslovni cilj je kreiranje prediktivnog modela mašinskog učenja, koji će otkrivene zakonitosti u podacima generalizovati i koristiti za predviđanje rizičnih investicija. Ovaj model će se trenirati na istorijskim podacima kako bi naučio obrasce i korelacije između različitih karakteristika vozila i njihovih ishoda na tržištu. Nakon što se model obuči i validira, koristiće se za procenu rizika novih vozila na aukcijama, što će omogućiti dilerima da unapred identifikuju potencijalno problematične kupovine. Na ovaj način, model će doprineti smanjenju finansijskih gubitaka i optimizaciji poslovanja.

Očekujemo izazove u identifikaciji ključnih faktora koji utiču na kvalitet predikcija, s obzirom na kompleksnost podataka i potencijalne nedoslednosti u njima. Takođe, značajan izazov biće obezbeđivanje da model generalizuje otkrivene obrasce na nove podatke, kako bismo izbegli rizik od loših investicija u realnim poslovnim uslovima.

Efekat koji želimo da postignemo je model koji na stvarnim podacima donosi takve odluke da stvara veću finansijsku dobit od trenutne. Benčmark će biti model koji ne pokušava da prepozna anomaliju, već uvek “kupuje” vozilo . Cene grešaka date su u tabeli. Od ostalih metrika , veći značaj ćemo dati odzivu, jer nam je skuplja greška kada se kupi loš automobil nego kada se kupi dobar - kupiće se neki drugi dobar. Modele ćemo takođe porediti po AUC-u.

*Tabela 1 Pregled različitih scenarija kupovine vozila na aukciji i njihovog uticaja na konačni prihod, uzimajući u obzir predikciju modela i stvarno stanje vozila.*

| Predikcija | Ishod | Na aukciji | Prodajna cena | Računovodsveni prihod | Ekonomski prihod |
|---|---|---|---|---|---|
| Ne kupiti | Loša kupovina | Ne kupiti: $0 | - | $0 | $0 |
| Kupiti | Dobra kupovina | Kupiti: -$15,500 | $18,600 | $3,100 | $3,100 |
| Ne kupiti | Dobra kupovina | Ne kupiti: $0 | - | $0 | -3,100$ |
| Kupiti | Loša kupovina | Kupiti: -$15,500 | $9,200 | -$6,300 | -$6,300 |

## Razumevanje I priprema podataka

Podatke smo preuzeli sa takmičenja ” Don’t Get Kicked!”, u organizaciji kompanije Karvana, na sajtu *Kaggle.com*[1]. Skup podataka ima 32 atributa I binarnu izlaznu promenljivu, koja označava da li je automobile *kick* ili ne i 72 983 instanci. Podatke smo podelili na trening, validacioni i test skup u odnosu 70-15-15. Analizu podataka sproveli smo na trening skupu, dok smo skaliranje, standardizaciju i popunjavanje nedostajućih vrednosti primenili na svim skupovima koristeći isključivo statistike (poput proseka i standardne devijacije) izračunate na trening skupu, kako bismo izbegli curenje informacija između skupova.

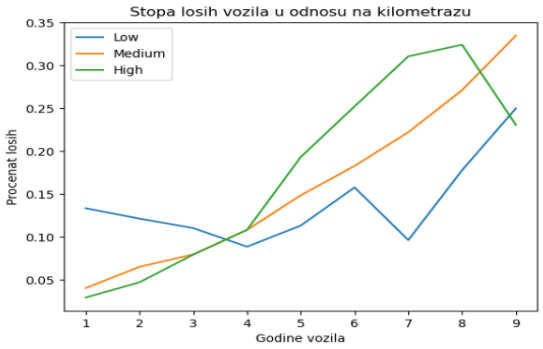
Neki od glavnih atributa su osnovna cena vozila (‘VehBCost’), kilometraža (‘VehOdo’), indikatori cena automobila (‘Manheim Market Report’), starost vozila i drugi. Podaci su neizbalansirani po izlaznoj promenljivoj , 12.3% čine pozitivne instance (“Is Bad Buy” = 1).

Prvo značajno zapažanje je uticaj starosti vozila na starosti vozila na stopu loših. Na grafikonu možemo videti da zastupljenost loših vozila raste po grupama starosti (4% vozila starosti od jedne godine, do 33% vozila starosti od 9 godina).

Prodaju se vozila od 33 različita proizvođača, četiri najveća proizvođača su Ševrolet, Dodž, Ford I Krajlser, koji su zajedno proizveli 70% automobila iz skupa, a 84% vozila je američkog porekla. Samim tim ne čudi podatak da je 96% vozila ima automatski tip menjača. Karvana najviše posluje na jugu I jugozapadu SAD-a, preko 50% kupovina obavilo u svega 4 savezne drzave - Teksas, Florida, Kalifornija i Severna Karolina.

Jedan od podataka koji nas je interesovao bio je komplet opreme (Trim). Svaki proizvođač ima svoju oznaku za nivo opreme, a često I variraju od modela do modela. Ukupno je bilo 134 različite oznake. Sve njih smo istražili I kategorisali u kategorije *Basic, Mid-Range, Sport, Luxury, Special.* Potom smo uradili Hi-kvadrat da proverimo značajnost korelacije sa izlaznom promenljivom, I zaključili da postoji značajna statistička veza između ovih kategorija. Nedostajuće vrednosti za Trim smo popunili na osnovu modela vozila, koji je bio poznat za sve instance.

Inicijalna pretpostavka je bila da će kilometraža biti ključna za predviđanje. Vozila u prvom kvantilu po pređenim kilometrima su bila loša u 8.7% posto slučajeva, a u poslednjem kvantilu u 16.2%. Da bismo dalje utvrdili uticaj kilometraže, grupisali smo vozila po godinama starosti (od 1 do 9) I svaku grupu smo dalje podelili na tri grupe (*low, medium,* high) , kako bismo otkrili vozila koja su puno prešla u odnosu na svoju starost. Rezultati su jasno potvrdili našu pretposavku.




*Grafik 1 Stopa loših vozila u odnosu na kilometražu i godine*

[1] https://www.kaggle.com/c/DontGetKicked

Kilometražu smo standardizovali koristeći prosek I standardnu devijaciju posebno za svaku godinu.

Imali smo osam MMR indeksa, četiri od njih označavaju očekivane cene vozila u trenutku kupovine u zavisnosti od stanja vozila I tršišta (prosečno I dobro stanje, kupovina na aukcijama I dilerima). Preostali MMR indeksi se tiču ovih cena u trenutku kreiranja seta. Hipotetički, ako bismo napravili model koji bi se koristio u industriji, ovi podaci ne bi bili relevantni, zato što bi vremenom postali zastareli jer se vezuju za jedan trenutak u vremenu. Takođe, pojavljivaće se novi modeli automobila, za koje ne postoje cene iz tog trenutka. Zato ih nećemo koristiti u modelu. Prva četiri smo sveli u jedan, njihov prosek, jer imaju izrazito visok stepen linearne zavisnosti.

Pored navedenih, izveli smo još tri nova atributa – cena osiguranja po kilometru, osnovna cena vozila po kilometru I cena po starosti.

Sve numeričke atribute smo standardizovali, a kategoričke smo pretvorili u više binarnih, metodom *OneHotEncoding*.

Sproveli smo klasterovanje *K-Means* algoritmom, lakat krivom smo se opredelili za tri klastera I dodali smo atribut koji označava pripadnost klasteru, što se pokazalo kao korisno u razumevanju pravilnosti u podacima, ali I klasifikaciji. Takođe, zbog neizbalansiranosti klasa, primenili smo tehnike resamplinga. Primenili smo undersampling, čime smo smanjili broj negativnih instanci, I oversampling (tehniku *SMOTE*), čime smo veštački generisali pozitivne instance. Ovo je dovelo do velikog pada klasifikacionih modela na stvarnim podacima, pa nismo ozbiljnije razmatrali taj novi skup.

Tokom modelovanja smo generisali nove atribute polinomijalnom transformacijom kako bismo uveli nelinearnost I sproveli smo analizu značajnosti atributa različitim tehnikama – Gini, Lasso, Permutation Importance. Koristili smo Laso tehniku za odabir atributa, među kojima su bili I polinomijani. Najznačajni atributi bili su VehOdoStandard, VehicleAge, MMRAverage, WarrantyPerMile.

**Modeliranje I evaluacija**

U ovoj fazi smo obučili tri modela mašinskog učenja. Sve modele ćemo obučaavati na trening setu kros validacijom, zatim ćemo podesiti hiperparametre na validacionom setu koristeći GridSearch. Odabraćemo najbolji model od dobijenih, njega obučiti na spojenom trening I validacionom setu I tesirati na test setu. Na taj način ćemo izbeći bilo kakvo curenje podataka I dobićemo što realniju ocenu modela. Potom ćemo izračunati prag odlučivanja koji bi nam doneo najviše finansijske dobiti I uporedićemo ga sa finansijskom dobiit koju bismo imali da smo kupili sve automobile iz test seta.

Odlučili smo se za logističku regresiju, Random Forest I XGBoost. Logistička regresija je jednostavan I interpretabilan model I na njoj smo doneli sledeće odluke – kako tretirati klastere – kao atribut ili kreirati model za svaki klaster I koju tehniku odabira atributa koristiti. Model za svaki kluster nije dala značajne rezulate, pa smo zbog jednostavnosti izabrali da pripadnost klasteru koristimo kao atribut. Za selekciju atributa kod svih modela smo koristili laso. Random Forest je ansambl model koji kombinuje predikcije više stabala odlučivanja. Njegova struktura čini ga otpornijim na *overfitting* u poređenju sa pojedinačnim stablima odlučivanja, posebno u situacijama sa mnogo varijabli, a takođe su I fleksibilna po pitanju hiperparametara. Na Random Forestu smo doneli odluku o resamplovanju podataka. Kako smo resamplovali samo trening set, modeli nisu bili u stanju da detektuju pozitivne instance, nisu bili dovoljno kompleksni. Balansiranje smo

postigli hiperparametrom *class_weight*. XGBoost je poznat po svojoj sposobnosti da postigne visoku tačnost u predikcijama. To je jedan od najsnažnijih modela za klasifikaciju i često se koristi u situacijama kada je potrebno izvući maksimalne performanse iz podataka.

| Model | Tačnost | Preciznost | Odziv | F1 statistika | AUC | Ekonomska dobit |
|---|---|---|---|---|---|---|
| Logistička regresija | 0.878 | 0.5116 | 0.0165 | 0.03 | 0.6731 | --- |
| Random Forest | 0.6154 | 0.1895 | 0.6564 | 0.2941 | 0.6844 | $21,390,300 |
| XGBoost | 0.6563 | 0.1958 | 0.5846 | 0.2933 | 0.6786 | $21,399,000 |
| Benčmark | | | | | | $21,377,300 |

*Tabela 2 Rezultati na validacionom setu najboljih modela*

Logistička regresija je dala ubedljivo najslabije rezultate, iako su sve vrednosti AUC-a blizu. Odziv regresije je na svega 1%, što znači da model ni ne pokušava da predvidi pozitivnu klasu. Relativno visok AUC se objašnava veoma niskim pragom odlučivanja. Random Forest I XGBoost su dali jako slične rezultate, oba su prebacila benčmark vrednost ali je XGBoost doneo veći prihod za 9,000 dolara.

## Implementacija

XGBoost je model koji ćemo koristiti u fazi implementacije.

| XGBoost na test setu | |
|---|---|
| Tačnost | 0.6588 |
| Preciznost | 0.2027 |
| Odziv | 0.6043 |
| F1 – statistika | 0.6844 |
| AUC | 0.3036 |
| Finansijska dobit | $21,327,900 |
| U odnosu na benčmark | +$50,900 |

*Tabela 3 Rezulzati na test setu*

## Zaključak I ideje za dalji rad

Tokom projekta uspeli smo da identifikujemo važne zakonitosti u podacima koje značajno utiču na odluku o tome da li je vozilo rizična investicija. Ključni faktori koje smo otkrili uključuju starost vozila, odnos starosti i pređene kilometraže, kao i korelaciju između cene sa osiguranjem, cene bez osiguranja i kilometraže. Kreirali smo prediktivni model koji zadovoljava inicijalne ciljeve postavljene na početku projekta, kao što su identifikacija rizičnih investicija i povećanje finansijske dobiti. Iako smo postigli prihvatljive rezultate u detekciji važnih faktora i razvoju prediktivnog modela, evidentno je da model nije bio dovoljno kompleksan, a izvedeni atributi nisu bili dovoljno informativni. U fazi modelovanja nedostajalo je tehničkog znanja i vremena za testiranje većeg broja modela, tehnika i pristupa, što je moglo značajno poboljšati konačne rezultate. Generalno, zadovoljni smo dobijenim rezultatima i smatramo da smo značajno produbili naše praktično i teorijsko znanje. Za buduće radove, preporučujemo da se više istraži detekcija anomalija, složenije tehnike klasterovanja, resamplovanje podataka i, naravno, primena neuronskih mreža, kako bi model bio spremniji za svakodnevnu upotrebu u realnim poslovnim uslovima.