Analiza preživljavanja na Titaniku: Primena klasifikacionih algoritama na skupu podataka putnika

Jelena Adamović, SV 6/2021

Definicija problema

Ovaj projektni zadatak predstavlja sveobuhvatnu analizu podataka o preživelima na Titaniku kroz primenu raznovrsnih algoritama klasifikacije, fokusirajući se na klasifikaciju u dve osnovne kategorije: putnici koji su preživeli i oni koji, nažalost, nisu. Cilj istraživanja je dubinsko razumevanje faktora koji su imali presudan uticaj na preživljavanje putnika tokom katastrofe Titanika.

Motivacija

Rezultati istraživanja pružaju uvid u važnost različitih faktora kao što su pol, klasa, starost i prisustvo članova porodice u predviđanju preživljavanja putnika na Titaniku. Takođe, ovi rezultati mogu biti korisni za dalja istraživanja u oblasti analize preživljavanja u sličnim katastrofalnim događajima.

Skup podataka

Skup podataka sadrži sledeće kolone: PassengerId, Survived, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked. **PassengerId**: Ova kolona sadrži jedinstveni identifikator svakog putnika na Titaniku. Svaki putnik ima svoj jedinstveni broj koji omogućava precizno praćenje i identifikaciju putnika u skupu podataka. **Survived**: Ova kolona označava da li je putnik preživeo ili nije preživeo. Vrednost 1 označava da je putnik preživeo, dok vrednost 0 označava da putnik nije preživeo. **Name**: Ova kolona sadrži imena putnika. **Sex**: Kolona "Sex" označava pol putnika, pri čemu vrednosti mogu biti "male" za muški pol ili "female" za ženski pol. **Age**: Ova kolona sadrži godine starosti putnika na Titaniku. **SibSp**: Ova kolona označava broj braće/sestara ili supružnika koje je putnik imao na brodu. **Parch**: Kolona "Parch" označava broj roditelja ili dece koje je putnik imao na brodu. **Ticket:** Ova kolona sadrži informacije o broju karte putnika. **Fare**: Ova kolona sadrži cenu karte koju je putnik platio za putovanje na Titaniku. **Cabin**: Kolona "Cabin" sadrži informacije o kabini u kojoj je putnik boravio tokom plovidbe Titanikom. Ova informacija može biti korisna u analizi preživljavanja, jer su određene lokacije na brodu možda bile povezane sa većom ili manjom verovatnoćom preživljavanja. **Embarked**: Ova kolona označava luku u kojoj je putnik ušao na brod (C = Cherbourg, Q = Queenstown, S = Southampton).

Trening skup bi bio podskup originalnog skupa podataka o putnicima Titanika koji bi se koristio za obuku modela. Ovaj skup bi sadržao podatke o putnicima kao što su pol, klasa, starost, prisustvo članova porodice i slično. Model bi se trenirao na ovim podacima kako bi naučio kako da predviđa ishod preživljavanja (Survived).

Broj uzoraka u train skupu je 891 dok je broj uzoraka u test skupu 418.

Ciljno obeležje je Survived i njegove vrednosti mogu biti 0 (nije preživeo) i 1 (preživeo je).

Link do dataset-a koji će biti iskorišćen za izradu projektnog zadatka je: Titanic dataset

Način pretprocesiranja podataka

U preprocesiranju podataka, pristupamo obradi skupa podataka kako bismo ih pripremili za dalju analizu i modeliranje. U ovom planu, prvo se fokusiramo na uklanjanje nedostajućih vrednosti. Kolona "Age" ima oko 20% nedostajućih vrednosti, što može značajno uticati na analizu. Kako bismo popunili ove nedostajuće vrednosti, koristimo pristup koji uzima u obzir klasu u kojoj je putnik putovao. Ovaj metod odabira prosečne vrednosti godina za svaku klasu, omogućavajući nam da očuvamo važne informacije o godinama putnika.

Drugo, kolona "Cabin" ima veliki broj nedostajućih vrednosti. Budući da nedostajuće vrednosti čine značajan deo ove kolone, odlučujemo da je potpuno izbacimo iz analize kako bismo izbegli izobličenje rezultata.

Nakon toga, primenjujemo tehnike enkodiranja kako bismo se nosili sa kategoričkim atributima. Koristimo one-hot encoding za kolonu "Embarked", koja označava luku u kojoj se putnik ukrcao. Ova tehnika transformiše kategoričke podatke u binarne vektore, čime se omogućava njihova efikasna obrada od strane algoritama mašinskog učenja.

Za kolonu "Sex", koristimo label encoding. Ova tehnika dodeljuje jedinstvene numeričke vrednosti svakoj kategoriji, što omogućava algoritmima da bolje razumeju i koriste ove informacije.

Metodologija

Prikupljanje podataka za ovaj istraživački projekat sprovedeno je putem Kaggle platforme, gde smo preuzeli odgovarajući dataset koji sadrži informacije o putnicima Titanika. Nakon prikupljanja podataka, sledeći korak je bio preprocesiranje podataka kako bismo ih pripremili za analizu i modeliranje.

U preprocesiranju podataka, primenili smo niz koraka kako bismo očistili podatke i pripremili ih za dalju analizu. Prvo smo identifikovali nedostajuće vrednosti u kolonama "Age" i "Cabin" i odlučili da uklonimo kolonu "Cabin" zbog velikog broja nedostajućih vrednosti. Nedostajuće vrednosti u koloni "Age" popunili smo prosečnim vrednostima godina za svaku klasu putnika, koristeći klasu kao faktor za procenu prosečne starosti.

Nakon toga, primenili smo tehnike enkodiranja kako bismo se nosili sa kategoričkim atributima. Za kolonu "Embarked" koristili smo one-hot encoding, dok smo za kolonu "Sex" koristili label encoding. Ovo je omogućilo pretvaranje kategoričkih atributa u numeričke oblike, što je neophodno za rad algoritama mašinskog učenja.

Kada su podaci bili pripremljeni, pristupili smo izboru modela za klasifikaciju. Odabrali smo četiri različita modela: Logističku regresiju, Naivni Bayes, SVM (Support Vector Machine) i Neuronsku mrežu za klasifikaciju. Svaki od ovih modela ima svoje prednosti i mane, te smo ih odabrali kako bismo dobili raznolike perspektive u analizi podataka o preživljavanju na Titaniku.

Nakon izbora modela, pristupili smo podešavanju hiperparametara modela kako bismo optimizovali njihove performanse. Korišćenjem trening i test skupa, trenirali smo svaki model i evaluirali njegove performanse. Evaluacija performansi obuhvatila je analizu tačnosti, preciznosti, odziva i F1-mere za svaki model.

Konačno, nakon evaluacije performansi, pristupili smo reviziji rezultata i potencijalnom podešavanju modela i parametara kako bismo poboljšali njihove performanse. Ovaj iterativni proces omogućio nam je

da donesemo pouzdane zaključke o efikasnosti različitih modela u predviđanju preživljavanja putnika na Titaniku.

Način evaluacije

Kada je reč o evaluaciji rezultata, koristimo se nizom metrika kako bismo detaljno procenili performanse modela. Glavne metrike koje koristimo za evaluaciju modela uključuju:

- Confusion Matrix (Matrica konfuzije): Pomoću matrice konfuzije možemo identifikovati broj tačnih predikcija za svaku klasu, kao i broj lažno pozitivnih i lažno negativnih predikcija.
- Accuracy (Tačnost): Ova metrika meri ukupnu tačnost modela, odnosno procenat tačnih predikcija u odnosu na ukupan broj predikcija.
- Sensitivity (Osetljivost): Ova metrika meri procenat pravilno klasifikovanih pozitivnih instanci u odnosu na ukupan broj stvarnih pozitivnih instanci.
- Specificity (Specifičnost): Ova metrika meri procenat pravilno klasifikovanih negativnih instanci u odnosu na ukupan broj stvarnih negativnih instanci.
- Recall (Recall): Ova metrika je sinonim za osetljivost i meri procenat stvarnih pozitivnih instanci koje su ispravno identifikovane od strane modela.
- F1 Score (F1 Skor): Ova metrika predstavlja harmonijsku sredinu između preciznosti i odziva.

Tehnologije

Za realizaciju ovog problema klasifikacije preživljavanja na Titaniku, koristili smo sledeće tehnologije i programske jezike:

- Python je osnovni programski jezik koji smo koristili za implementaciju celokupnog projektnog zadatka.
- TensorFlow je biblioteka otvorenog koda za mašinsko učenje koju smo koristili za implementaciju neuronske mreže za klasifikaciju.
- scikit-learn je biblioteka otvorenog koda za mašinsko učenje koja pruža širok spektar algoritama za mašinsko učenje.

Relevantna literatura

1. Joldžić O., Kosić D., *Mašinsko učenje*, Akademska misao, Beograd, 2020.
2. Geron A., *Mašinsko učenje: Scikit-Learn, Keras i TensorFlow; Koncepti, alati i tehnike za izgradnju inteligentnih sistema*, Mikro knjiga, Beograd, 2021.