Predikcija cene polovnog automobila na osnovu oglasa

Jelena Kovač E2 14/2024, Boško Kulušić E2 24/2024

1. Definicija problema

Predikcija cene polovnog automobila koristeći podatke iz oglasa. Na osnovu karakteristika automobila, vršiće se predikcija njegove cene. Cilj ovog projekta je procena tržišne vrednosti cene automobila kao i pomoć zajednici prilikom kupovine ili prodaje automobila.

2. Motivacija problema rešavanog u projektu

Određivanje da li je cena za prodaju ili kupovinu polovnog automobila odgovarajuća predstavlja jednu od prepreka sa kojom se suočava zajednica prilikom kupovine ili prodaje automobila. Takođe, tržište automobila je industrija u stalnom usponu, čija se vrednost skoro udvostručila u poslednjih nekoliko godina. Različiti sajtovi koriste različite algoritme za predviđanje cene, stoga ne postoji jedinstven algoritam za predviđanje cene.

3. Relevantna literatura

3.1. Application of Machine Learning Techniques to Predict the Price of Pre-Owned Cars in Bangladesh, 2021 [pdf]

Tema rada: Procena cene polovnog automobila na tržištu Bangladeša upotrebom tehnika mašinskog učenja. Cilj rada im je bio da naprave model koji će da iskoriste u web aplikaciji.

Podaci: Autori su kreirali skup podataka koristeći tehnike web sceaping-a sajta za onlajn prodaju u Bangladešu bikroy.com. Taj skup sadrži 1209 podataka sa 10 obeležja (ime, marka, model, godište, menjač, tip karoserije, gorivo, snaga motora, kilometraža, i ciljno obeležje cena izražena u valuti bangladeška taka). Obeležje ime je odbačeno zbog previše jedinstvenih vrednosti.

Metodologija: Podaci su pretprocesuirani (čišćenje, normalizacija) i iskoristili su tehnike poput label encoding-a i pretvaranja kategoričkih obeležja u binarne kolone. Upotrebljeno je 5 modela mašinskog učenja: linearna regresija, LASSO regresija, stablo odlučivanja (decision tree), Random Forest, i extreme gradient boosting (XGBoost). Ulaz u modele su obeležja automobila dok je izlaz prediktovana cena.

Evaluacija i rezultati: Skup je podeljen na trening i test u odnosu 80:20. Korišteni su srednja apsolutna greska (engl. *Mean Absolute Error*, **MAE**), koren srednja kvadratna greška (engl. *Root Mean Squared Error*, **RMSE**) i R-kvadrat (engl. *R-squared*, $\mathbf{R^2}$). Zbog velikih vrednosti za tumačenje su koristili log vrednosti RMSE i MAE (log MAE i log RMSE). XGBoost je pokazao

najveće vrednosti R2 (91%). XGBoost i Random forest su ostvaili najbolji rezultat posmatrajući log MAE (11.72% i 11.63%) i log RMSE (12.53% i 12.6%).

Zaključak: XGBoost i Random forest bi mogli dati dobre rezultate i na našem skupu podataka te ćemo ih koristiti u radu. Iz ovog rada ćemo iskoristiti i web scraping za prikupljanje podataka, tehniku label encoding i podelu na trening i test skup. Razlika je što će naši podaci uključivati više obeležja.

3.2. A Deep Learning Approach for Used Car Price Prediction, 2022 [pdf]

Tema rada: Procena da li ANN može precizno da predvidi cenu polovnih automobila. Rezultati se upoređuju sa tradicionalnim regresionim tehnikama.

Podaci: Podaci su preuzeti sa sajta keggle.com. Nakon obrade, 140,000 vozila je iskorišteno za trenig a 30,000 za testiranje. Podaci se nalaze na linku. Podaci uključuju obeležja kao što su: model, brend, boja, godište, pređena kilometraža, menjač, gorivo, stanje vozila, regija, država, i ciljno obeležje cena.

Metodologija: Pored pretprocesiranja podataka, autori su razvili feed-forward ANN. Takođe, koristili su tradicionalne regresione tehnike: linearnu, Decision Tree, Random Forest, Gradient Boosting. Ulaz u modele predstavljaju obeležja automobila, dok je izlaz prediktovana cena.

Evaluacija i rezultati: Pored trening i test skupa, 20% testing skupa je preuzeto za validacioni skup. Korištene su **MAE**, **MAPE** (engl. *Mean Absolute Percentage Error*), **RMSE** i $\mathbf{R^2}$. ANN model je postigao najbolji učinak prema svim metrikama sa MAPE od 0.11, RMSE od 2104. Random Forest Regressor je ostvario drugi najbolji rezultat sa MAPE od 0.14, RMSE od 2717.79. Linearni regresor je ostvario najlošije rezultate.

Zaključak: U našem projektu ćemo koristiti Random Forest regresionu tehniku i ANN. Za evalaluacionu meru, koristićemo RMSE i MAPE. Najveća razlika je što naš skup podataka neće biti toliko velikog obima kao u ovom radu.

3.3. Used Car Price Prediction Model: A Machine Learning Approach, 2024 [pdf]

Tema rada: Primena tehnika mašinskog učenja za predikciju cene polovnih automobila. Korištene tehnike su KNN, linearna regresija i polinomijalna regresija.

Podaci: Podaci su prikupljeni tehnikom web scraping-a satja carsome.id. Podaci sadrže 504 reda i uključuju 13 obeležja kao što su: marka, garancija, lokacija, boja, godište, pređena kilometraža, lokacija, i ciljno obeležje cena izražena u indonežanskim rupijama.

Metodologija: Autori su izvršili eksplorativnu analizu podataka kao i njihovo pretporcesiranje. Korištene regresione tehnike: linerana regresija, polonomijalna regresija i KNN. Ulaz u modele predstavljaju obeležja automobila, dok je izlaz prediktovana cena.

Evaluacija i rezultati: Izvršena je podela skupa podataka na trening i test (odnos 70:30). Za meru evaluacije korištene su sledeće metode: **RMSE**, **MAPE** i **R^2** . KNN je ostvario najbolje rezultate po svim metrikama (RMSE : 36561034, MAPE : 0.083, R^2 : 0.988).

Zaključak: Slično kao u ovom radu, u našem projektu ćemo koristiti web scraping tehniku za prikupljanje podataka, podelu na trening i test skup, kao i evaluacionu metrike MAPE i RMSE.

4. Skup podataka

U našem projektu ćemo podatke prikupiti samostalno upotrebom web scraping-a sa sajta polovniautomobili.com i sadržaće oglase sa teritorije Beograda i Vojvodine za automobile između 2006. i 2015. godišta. Skup podataka će sadržati oko 2500 podataka.

Obeležja: Cena, Model, Marka, Starost, Kubikaža, Gorivo, Karoserija, Snaga motora, Kilometraža, Oštećenje, Pogon, Klima, Materijal enterijera, Menjač, Broj vrata, Boja, Emisiona klasa motora, Koliko poseduje sigurnosnih dodataka (Airbag, Child lock, ABS, Blokada motora, Ulazak bez ključa), Koliko poseduje napredne dodatne opreme (Sportska sedišta, Tempomat, Senzori za kišu, Parking senzori, Aluminijumske felne, Multimedija).

Ciljno obeležje je **cena** automobila u oglasu izražena u evrima. Kako su u pitanju stariji automobili, pretpostavljamo da će opseg cena da bude od 1.000 do 35.000 evra (moguća odstupanja jer nije kompletiran web scraping).

5. Metodologija

Nakon web scraping-a podataka, potrebno je izvršiti njihovo pretprocesiranje. Nedostajuće vrednosti koje se pojavljuju u kategoričkim obeležjima ćemo dopuniti novom kategoričkom vrednošću “nepoznato”, dok ćemo kontinualna dopuniti tehnikama poput popunjavanjem prosečne vrednosti, interpolacijom polinomom i slično. Ukoliko podatak sadrži više nedostajućih obeležja razmotrićemo njegovo uklanjanje. Detekcija autlajera će se vršiti pomoću DBScan algoritma nad cenom automobila. Normalizacija će biti primenjena nad obeležjima koja sadrže velike vrednosti (npr. cena, kubikaža). Kategorička obeležja malog broja kategorija (npr. menjač) će biti enkodirana pomoću one-hot enkodinga dok će za obeležja većeg broj kategorija (npr. boja) biti korišćen label encoding. Takođe, odradićemo i eksplorativnu analizu nad našim podacima u slučaju da su neka obeležja već sadražana u drugim ili nam ne donose nikakvu vrednost. Primer pitanja na koja ćemo odgovarati su: da li je skup podataka balansiran, da li neka obeležja prate isti trend pa je dovoljno koristiti jedno u predikciji, koliko boja ili određene performanse utiču na cenu, i slično. Ulaz u modele predstavljaju obeležja automobila, dok je izlaz prediktovana cena izražena u evrima. Za predikciju ćemo koristiti sledeće metode: Random Forest, XGBoost i feed-forward ANN jer su imali najbolje rezultate u refernciranim radovima.

6. Metod evaluacije

Skup podataka ćemo podeliti na trenig, test i validacioni skup (odnos 80:10:10). Za mere evaluacije ćemo koristiti MAPE i RMSE (za interpretaciju ćemo razmotriti upotrebu log RMSE).