Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 4.4
DOI: https://doi.org/10.24867/31BE15Topalovic

PREDIKCIJA CENA AVIONSKIH LETOVA UPOTREBOM ALGORITAMA MAŠINSKOG UČENJA

AIRLINE FLIGHT PRICE PREDICTION USING MACHINE LARNING ALGORITHMS

Saška Topalović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIČKO I RAČUNARSKO INŽENJERSTVO

Kratak sadržaj – *U ovom radu se istražuje problem predikcije cena letova. Predikcija cena avionskih letova predstavlja izazov kako za putnike, koji žele da plaćaju objektivnu cenu, tako i za avio-kompanije, koje nastoje da optimizuju prihode. Tradicionalne metode predikcije često ne uzimaju u obzir varijabilne faktore poput sezonskih oscilacija, potražnje i dostupnosti sedišta, što rezultuje neadekvatnim cenama. Stoga je razvijanje modela zasnovanog na mašinskom učenju važno kako bi se postigla veća transparentnost i efikasnost tržišta avionskih karata. Metodologija rada uključuje upotrebu sledećih algoritama mašinskog učenja: Random Forest, Decision Tree i XGBoost. Za evaluaciju modela su korišćeni trening i test skup podataka u odnosu 8:2. Evaluacijom performansi, pomoću RMSE (Root Mean Squared Error) metrike, Random Forest se pokazao kao najbolji model za predikciju cena avionskih karata.*

Ključne reči: *cena avionskog leta; mašinsko učenje; Random Forest; Decision Tree; XGBoost;*

Abstract – *This paper investigates the problem of flight price prediction. The prediction of flight prices poses a challenge for both passengers, who seek to pay a fair price, and airlines, which aim to optimize their revenues. Traditional prediction methods often fail to account for variable factors such as seasonal fluctuations, demand, and seat availability, resulting in inaccurate pricing. Therefore, developing a machine learning-based model is essential for achieving greater transparency and efficiency in the flight ticket market. The methodology employed includes the use of the following machine learning algorithms: Random Forest, Decision Tree, and XGBoost. The model evaluation was performed using an 8:2 train-test data split. Based on the performance evaluation using the RMSE (Root Mean Squared Error) metric, Random Forest proved to be the best model for flight price prediction..*

Keywords: *airline flight price; machine learning; Random Forest; Decision Tree; XGBoost;*

1. UVOD

Tržište avionskih karata je veoma dinamično i podložno brojnim promenama, što direktno utiče na cene avionskih letova, a samim tim i na putnike, avio-kompanije i druge relevantne aktere u industriji. Putnici nastoje da pronađu najpovoljnije avionske karte i tako optimizuju svoj budžet, a predikcija cena avionskih letova bi im omogućila da unapred planiraju i rezervišu svoja putovanja sa boljim razumevanjem očekivanih cena. Avio-kompanije bi, s druge strane, mogle da koriste predikciju cena letova kako bi bolje upravljale svojim tarifama, prilagodile ih potrebama potražnje i konkurencije, te optimizovale svoje prihode.

Razvoj modela za predikciju cena avionskih letova predstavlja veliki izazov zbog širokog spektra tehničkih i tržišnih faktora. Fokus ovog rada je implementacija algoritma mašinskog učenja koji će moći efikasno da analizira i obrađuje različite podatke kao što su karakteristike letova, istorijske cene i trendovi potražnje. Cilj je postići što veću preciznost u predikciji cena letova, uzimajući u obzir složenu i dinamičnu prirodu ovog tržišta.

Za kreiranje modela korišćena su tri algoritma mašinskog učenja: *Random Forest, Decision Tree* i *XGBoost*. Modeli su obučeni na podacima koji sadrže različite tehničke i tržišne karakteristike letova, kao što su vreme polaska, destinacija, avio-kompanija, broj presedanja, trajanje leta i istorijske cene. Evaluacija modela izvršena je podelom podataka na trening i test skupove u odnosu 8:2. Za merenje uspešnosti modela korišćena je RMSE (*Root Mean Squared Error*) metrika. Na osnovu dobijenih rezultata, *Random Forest* algoritam se istakao kao najprecizniji model za predviđanje cena letova, nadmašivši ostale modele po tačnosti.

U nastavku rada biće detaljno objašnjeni različiti aspekti rešavanog problema. U poglavlju 2 daje se pregled radova koji se bave sličnom tematikom. Poglavlje 3 sadrži opis skupa podataka, eksplorativnu analizu i pretprocesiranje podataka, kao i algoritme korišćene za predikciju cena letova. Poglavlje 4 je posvećeno diskusiji rezultata dobijenih primenom formiranih modela. Na kraju, poglavlje 5 sadrži zaključak koji rezimira ključne uvide.

2. SRODNA ISTRAŽIVANJA

U radu [1], Tziridis i saradnici su identifikovali skup karakteristika koje opisuju tipičan let, pretpostavljajući da one utiču na cenu. Studija se sastoji iz 4 faze. U prvoj fazi su eksperimentalno određena obeležja koja utiču na cenu primenom "*oneleave-out*" pravila. U drugoj fazi ručno su prikupljeni podaci sa veba, ukupno 1814 letova, gde je svaki let opisan sa 8 obeležja. U trećoj fazi odabrano je 8 modela mašinskog učenja i primenjeno na iste podatke. Na kraju, u četvrtoj fazi rešenje je evaluirano desetostrukom

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Aleksandar Kovačević, red. prof.

unakrsnom validacijom. Mera performansi korišćena za upoređivanje modela je MSE. Eksperimenti su pokazali da je uklanjanjem karakteristike „broj preostalih dana do poletanja“ rezultat bio najgori. Pored toga, *Bagging Regressor* i *Random Forest* su uvek imali dobre performanse. Rezultati su pokazali da su modeli mašinskog učenja zadovoljavajući za predviđanje cena avionskih letova i da dostižu tačnost čak do 88%.

Wang i saradnici predložili su novi okvir mašinskog učenja za predviđanje kvartalne prosečne cene avionskih karata na nivou tržišnog segmenta, fokusirajući se na određene kombinacije polazišta i odredišta [2]. Koristili su kombinaciju dva javno dostupna skupa podataka: *Airline Origin and Destination Survey* (DB1B) i *Air Carrier Statistics database* (T-100). Podaci su prikupljeni tokom 2018. godine. DB1B skup podataka obuhvata informacije o avio-liniji, dok T-100 skup podataka sadrži statističke podatke o avio-prevoznicima. Za evaluaciju modela koristili su RMSE i prilagođeni R kvadrat. Za kreiranje modela mašinskog učenja koristili su *Random Forest*, linearnu regresiju, SVM, *Multilayer Perceptrons* (MLPs) i *XGBoost Tree*. Zahvaljujući tehnikama selekcije obeležja, pokazalo se da je *Random Forest* model sposoban da predviđa kvartalnu prosečnu cenu avio-karte sa prilagođenim R kvadratom od 0.869.

3. METODOLOGIJA

U ovom poglavlju će biti predstavljena implementacija sistema za predikciju cena avionskih letova. Ulaz u sistem čini skup podataka o letovima, uključujući informacije o datumu polaska, avio-kompaniji, vremenu polaska i dolaska, broju presedanja i drugim faktorima. Izlaz sistema su predviđene cene karata na osnovu modela mašinskog učenja, koji su obučeni na prethodno pripremljenim podacima.

Poglavlje je organizovano u nekoliko potpoglavlja. Prvo će, u potpoglavlju 3.1, biti opisan skup podataka, uključujući njegove osnovne karakteristike. Zatim će, u potpoglavlju 3.2 biti opisana ekplorativna analiza i pretprocesiranje podataka. Dok će se poglavlje 3.3 fokusirati na obuku modela i optimizaciju hiperparametara.

3.1. Skup podataka

Skup podataka korišćen u radu je “*Flight Fare Prediction*”, javno dostupan, preuzet sa *Kaggle* platforme [3]. Sadrži informacije o avionskim letovima iz 2019. godine i sastoji se od ukupno 10683 zapisa. Svaki let u skupu podataka opisan je sa jedanaest obeležja, a to su: avio-kompanija (Airline), datum putovanja (Date_of_Journey), polazište (Source), odredište (Destination), ruta (Route), vreme polaska (Dep_Time), vreme dolaska leta (Arrival_Time), trajanje leta (Duration), ukupan broj presedanja tokom putovanja (Total_Stops), dodatne informacije o putovanju (Additional_Info), cena avionske karte (Price) Cena avionske karte predstavlja ciljno obeležje, tj. zavisnu promenljivu koja se predviđa na osnovu preostalih deset nezavisnih obeležja.

3.2. Eksplorativna analiza i pretprocesiranje podataka

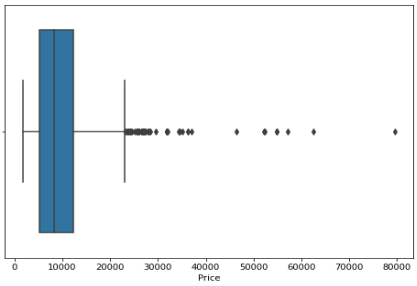
Najpre je izvršena analiza nedostajućih vrednosti, pri čemu je ustanovljeno prisustvo nedostajućih obeležja (NaN) u skupu podataka. Pošto je samo jedan let imao dva nedostajuća obeležja, on je uklonjen iz skupa podataka. Nakon toga, izvršena je identifikacija duplikata i ustanovljeno je da postoji 220 potpuno identičnih zapisa. Prisustvo više identičnih instanci ne doprinosi raznolikosti i kvalitetu podataka, već samo povećava nepotrebnu složenost. Da bi se rizik od grešaka i nepreciznosti sveo na minimum, ti duplikati su uklonjeni.

Zatim, neophodna je bila konverzija vremenskih obeležja. Promenljive *Date_of_Journey*, *Dep_Time*, *Arrival_Time* i *Duration* konvertovane su iz tipa *string* u tip *datetime*. Nakon ove konverzije, izvedena su nova obeležja dan (engl. *Day_of_Journey*) i mesec putovanja (engl. *Month_of_Journey*), sati (engl. *Dep_Hours*) i minuti (engl. *Dep_Minutes*) polaska, kao i vreme dolaska (engl. *Arrival_Hours* i *Arrival_Minutes*). Pored toga, izračunato je i ukupno vreme trajanja leta u minutama (engl. *Duration_Minutes*), što je omogućilo bolje razumevanje i analizu vremenskih obrazaca u podacima, čineći ih pogodnijim za dalje modelovanje.

Cena avionskih letova može varirati u zavisnosti od dana u sedmici, vremena polaska i meseca putovanja. Letovi vikendom i tokom turističke sezone ili praznika često imaju više cene, dok letovi kasno uveče ili noću, zbog manje potražnje, mogu biti jeftiniji. Analiza prosečnih cena pokazala je najpre da su cene letova vikendom neznatno više u odnosu na cene radnim danima, kao i da se nijedan dan u sedmici posebno ne izdvaja, ali su petkom i nedeljom cene blago više u odnosu na ostale dane. Pretpostavka da su noćni letovi (između 19h i 7h) u proseku nešto jeftiniji od dnevnih se pokazala tačnom. Najviša prosečna cena bila je u martu, a najniža u aprilu, prema podacima iz skupa koji obuhvataju period od marta do juna.

Uobičajena pojava u avio-industriji je da veći broj presedanja tokom putovanja obično rezultuje višim cenama avionskih karata, a analiza je to i potvrdila. Letovi sa četiri presedanja su značajno skuplji od direktnih letova.

Kako je cena jedino numeričko obeležje u skupu podataka, bilo je važno ispitati moguća odstupanja, odnosno *outlier*-e. U svrhu identifikacije *outlier*-a je korišćen *boxplot* grafikon, koji je pokazao da postoji šest cena koje se značajno razlikuju od ostalih (Slika 1).



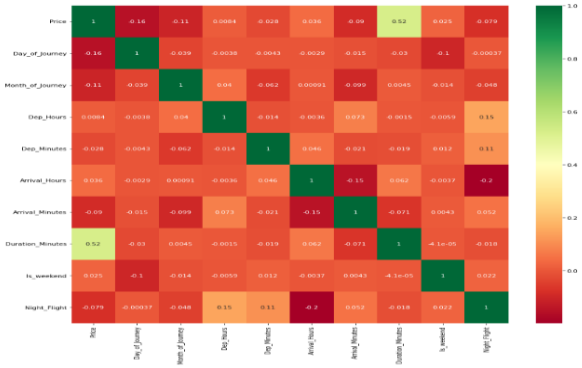
Slika 1. *Raspodela cena u skupu podataka*

Uočavanjem ovih *outlier*-a, sledeći korak je bio analiza prosečne cene letova po avio-kompanijama kako bi se utvrdilo da li neki od tih letova pripada *Business* klasi, što bi moglo objasniti njihovo odstupanje. Ispostavilo se da postoje letovi *Business* klase avio-kompanije *Jet Airways*, koji znatno odstupaju po ceni. Kako je u pitanju samo šest takvih letova, zaključeno je da, zbog malog broja primeraka, nije svrsishodno vršiti posebnu predikciju cena

za ovu klasu letova, pa su ti letovi uklonjeni iz skupa podataka.

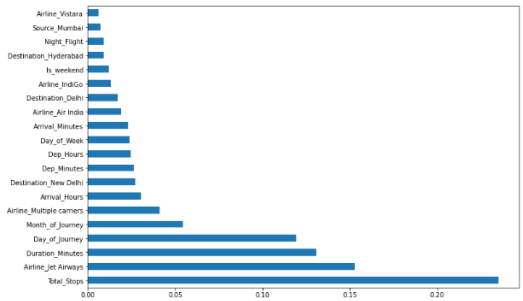
U cilju boljeg razumevanja raspodele kategoričkih obeležja, generisani su *barplot* dijagrami za svaku kategoriju. Posebno se izdvojio dijagram za promenljivu '*Additional_Info*', gde je oko 80% podataka pripadalo kategoriji '*No Info*'. Takođe, promenljiva '*Route*' pokazala je značajnu povezanost sa brojem presedanja, polazištem i odredištem putovanja, što smanjuje njenu nezavisnost. Kako ove dve promenljive ne doprinose informativnosti i raznolikosti, one su izuzete iz dalje analize.

Za analizu korelacije korišćen je *heatmap* kao vizuelni alat. Najviša korelacija, od 0.52, uočena je između cene avionskih letova i trajanja putovanja, što ukazuje na umerenu pozitivnu vezu između ove dve promenljive (Slika 2). To znači da duži letovi, u proseku, imaju više cene u odnosu na kraće letove. Ipak, važno je napomenuti da korelacija samo ukazuje na to da postoji neka vrsta povezanosti, ona ne implicira uzročno-posledičnu vezu između faktora. Ostala obeležja su pokazala značajno niže korelacije.



Slika 2. *Matrica korelacije skupa obeležja*

Za identifikaciju najznačajnijih obeležja u predviđanju cene letova korišćen je *Extra Tree Regressor*. Na osnovu rezultata prikazanih na Slika 3, utvrđeno je da je '*Total_Stops*', odnosno broj presedanja, najznačajnije obeležje sa najvećim uticajem na cenu leta.



Slika 3. *Vizualizacija 20 najvažnijih obeležja*

Važan korak pretprocesiranja je enkodovanje kategoričkih obeležja, budući da mnogi algoritmi mašinskog učenja zahtevaju da su ulazni podaci u numeričkom obliku. Za enkodovanje nominalnih kategoričkih obeležja (*Airline*, *Source* i *Destination*) korišćen je *one-hot encoding*. Broj presedanja (engl. *Total_Stops*) je jedino ordinalno kategoričko obeležje i za njegovo enkodovanje korišćen je *label encoding*, pri čemu su redni brojevi dodeljeni svakoj kategoriji, od '*non-stop*' (0) do najvećeg broja presedanja.

Na kraju, skup podataka je sortiran po datumu putovanja. Ovakvo sortiranje omogućilo je da se test skup formira kao vremenski interval nakon trening skupa, čime je simulirano predviđanje budućih cena na osnovu istorijskih podataka, što dodatno povećava verodostojnost modela.

3.3. Obuka modela i optimizacija hiperparametara

Za potrebe obuke modela, odabrana su tri algortima mašinskog učenja *Random Forest*, *Decision Tree* i *XGBoost*. *Decision Tree* algoritam pravi predikcije kroz seriju binarnih odluka, postepeno deleći skup podataka na podskupove na osnovu karakteristika koje najviše doprinose rešenju problema. *Random Forest* obučava stotine stabala odluke na različitim uzorcima podataka, a konačni izlaz je prosek vrednosti svih modela ili odluka dobijena većinskim glasanjem, u zavisnosti od tipa problema (regresija ili klasifikacija). *XGBoost* primenjuje tehniku *boosting*-a, tj. integriše veliki broj slabih klasifikatora kako bi formirao snažan model.

Svaki od pomenutih algoritama obučen je na pretprocesiranim podacima korišćenjem *test-train* podele u odnosu 8:2. Mera korišćena za evaluaciju performansi modela je RMSE (*Root Mean Squared Error*).

Nakon inicijalne obuke, izvršena je optimizacija hiperparametara. U tu svrhu korišćena je tehnika *RandomizedSearchCV*, koja omogućava ispitivanje različitih kombinacija hiperparametara unutar unapred definisanih opsega. Ovaj pristup ispituje nasumično izabrane kombinacije, čime se značajno ubrzava proces optimizacije u poređenju sa klasičnim *GridSeachCV*-om. Za svaki model, proces optimizacije hiperparametara obuhvatao je 10 iteracija slučajnog pretraživanja. Za svaku kombinaciju hiperparametara primenjen je kros-validacioni proces sa 5 podela (engl. *5-fold cross-validation*), gde je glavna metrika za ocenjivanje modela bila RMSE.

4. REZULTATI I DISKUSIJA

Prethodno obučeni modeli su testirani na skupu od 2092 leta koja nisu bila uključena u trening skup. Rezultati, prikazani u TABELA I, pokazuju performanse sva tri modela pre i posle optimizacije hiperparametara, merene vrednostima RMSE.

TABELA I. PERFORMANSE MODELA PRE I POSLE OPTIMIZACIJE HIPERPARAMETARA

Model	**Pre optimizacije hiperparametara**	**Posle optimizacijom hiperparametara**
	RMSE	***RMSE***
Random Forest	2042.709	1899.209
Decision Tree	3345.024	2505.235
XGBoost	1908.776	1971.674

Pre optimizacije, *XGBoost* model je postigao najnižu RMSE vrednost od 1908.776, što ga čini najboljim modelom u početnim uslovima. Nakon optimizacije hiperparametara, *Random Forest* model je zabeležio poboljšanje, smanjivši RMSE na 1899.209, što ga čini najboljim modelom nakon optimizacije. Ovaj rezultat nije iznenađujući, s obzirom na to da se u radu [2] takođe pokazao kao model sa najboljim performansama,

nadmašivši *XGBoost*. Slično tome, u radu [1] su eksperimenti pokazali da *Random Forest* konzistentno pokazuje dobre rezultate u predikciji cena letova, što ukazuje na njegovu pouzdanost i stabilnost u različitim kontekstima. S druge strane, *XGBoost* model, koji je prvobitno imao najbolje rezultate, pokazao je blagi porast RMSE, ali je i dalje zadržao visoke performanse, dok se *Decision Tree* pokazao kao najlošiji model za rešavani problem.

Blagi pad rezultata nakon optimizacije hiperparametara, za *XGBoost* model je bio iznenađujući, međutim može se objasniti složenošću prostora hiperparametara koji se koriste, kao i brojem parametara koji su razmatrani tokom pretrage. *RandomizedSearchCV* algoritam je, zbog ograničenog broja iteracija, možda propustio ključne kombinacije koje bi dale bolje rezultate. Pored toga, nasumičnost algoritma može dovesti do neravnomernog istraživanja prostora parametara, što potencijalno znači da određene oblasti nisu bile adekvatno istražene. Ovo ukazuje na potrebu za detaljnijom pretragom prostora hiperparametara, kako bi se iskoristio pun potencijal *XGBoost* modela.

Analizom grešaka uočeno je da su sva tri modela ostvarila najveće greške u predikciji letova sa visokim cenama. Detaljnijom analizom otkriveno je da modeli imaju tendenciju da prave veće greške prilikom predviđanja cena za letove sa visokim cenama, ali relativno kratkim trajanjem. Ovaj fenomen može biti posledica uočene umerene pozitivne korelacije između cene avio-karte i trajanja leta, koja je detaljnije obrađena u poglavlju 3.2. Korelacija ukazuje na to da duži letovi obično imaju više cene, dok su kraći letovi često jeftiniji. Međutim, ove anomalije sugerišu da postoje specifični letovi gde visoke cene nisu u skladu sa očekivanjima na osnovu trajanja leta, što rezultuje većim greškama modela.

5. ZAKLJUČAK

U ovom radu predstavljen je sistem za predikciju cena avionskih letova. Putnici se često oslanjaju na procene cena letova koje pružaju turističke agencije ili veb portali za rezervaciju, ali te procene ne odražavaju uvek realne tržišne trendove. Takođe, zanemaruje se činjenica da avio-kompanije često prilagođavaju tarife različitim tržišnim faktorima. Razvoj modela za objektivnu procenu cena letova omogućio bi putnicima da plaćaju fer cenu, dok bi avio-kompanije mogle bolje prilagoditi svoje tarife tržišnim uslovima, čime bi poboljšale svoje poslovne performanse.

Primenom tehnika mašinskog učenja, razvijeni su modeli za predikciju cena letova korišćenjem algoritama *Random Forest*, *Decision Tree* i *XGBoost*. Nakon optimizacije hiperparametara, performanse modela su evaluirane metrikom RMSE, koja je izabrana zbog svoje osetljivosti na velike greške. Najbolje rezultate postigao je model *Random Forest*, sa RMSE vrednošću od 1899.209 na test podacima. Međutim, analiza grešaka pokazala je da modeli uglavnom greše pri predviđanju letova sa visokim cenama i kraćim trajanjem.

Ovaj rad doprinosi boljem razumevanju faktora koji utiču na cenu letova i postavlja osnovu za dalja istraživanja u oblasti predikcije cena avionskih letova. Dalji pravci istraživanja i potencijalna poboljšanja sistema mogli bi obuhvatiti primenu naprednijih tehnika optimizacije hiperparametara, kao što su *Bayesian Optimization* ili *GridSearchCV* sa većim brojem iteracija, kako bi se istražile sve ključne kombinacije. Takođe, predikcija cena je trenutno ograničena na korišćeni skup podataka, te bi razmatranje proširivanja skupa dodatnim relevantnim faktorima, poput broja dana do polaska, klase leta, pozicije sedišta u avionu ili broja besplatnog prtljaga, moglo doprineti tačnosti modela. Osim toga, analiza grešaka ukazuje na potrebu za prilagođavanjem modela u smislu boljeg tretiranja letova sa kraćim trajanjem i višim cenama, možda kroz uvođenje dodatnih relevantnih karakteristika ili kroz unapređenje algoritama koji su osetljiviji na ove specifičnosti.

6. LITERATURA

[1] Tziridis, K., Kalampokas, T., Papakostas, G. A., & Diamantaras, K. I. (2017, August). Airfare prices prediction using machine learning techniques. In *2017 25th European Signal Processing Conference (EUSIPCO)* (pp. 1036-1039). IEEE.

[2] Wang, T., Pouyanfar, S., Tian, H., Tao, Y., Alonso, M., Luis, S., & Chen, S. C. (2019, July). A framework for airfare price prediction: a machine learning approach. In *2019 IEEE 20th international conference on information reuse and integration for data science (IRI)* (pp. 200-207). IEEE.

[3] https://www.kaggle.com/datasets/nikhilmittal/flight-fare-prediction-mh (pristupljeno u martu 2024.)

Kratka biografija:

Saška Topalović je rođena 31.08.2000. godine u Doboju, gde je stekla svoje osnovno i srednje obrazovanje. Školske 2019/20. godine se upisuje na Fakultet tehničkih nauka u Novom Sadu na studijski program softversko inženjerstvo i informacione tehnologije. Diplomski rad pod nazivom „Arhitektura nultog poverenja“ odbranila je 2023. Iste godine se upisuje na master akademske studije, na isti studijski program. 2024. je položila sve ispite predviđene planom i programom i stekla uslov za odbranu master rada.

kontakt: tsaska98@gmail.com