Predviđanje otkaza hotelskih rezervacija

Anđela Matijašević, Matko Soče, Vinko Bokšić

***Abstract*— Predviđanje otkaza rezeravacija u hotelskoj industriji područje je izraženog interesa. Cilj je razvijanje prediktivnog model metodama strojnog učenja kako bi se budući otkazi rezervacija mogli predvidjeti. U rješavanju problema primijenjene su metode: Decision tree, Random forest, Logistic regression i XGB classifier, a najbolji rezultat daje Random forest.**

Ključne riječi—strojno učenje, Decision tree, Random forest, Logistic regression, XGB classifier, previđanje otkaza rezevacija

I. UVOD

Uspješno realizirane rezervacije i što veća popunjenost kapaciteta primarni su ciljevi svakog hotelskog menadžmenta. U današnje vrijeme zbog visoke konkurentnosti hoteli nude otkaze rezervacija s potpunim povratom novca, čime gostima omogućuju fleksibilnost do zadnjeg trenutka, dok istodobno sebe izlažu riziku gubitka prihoda. Budući da hoteli žele ostati poželjni potencijalnim gostima razvijanje strogih politika otkaza plaćanja u potpunosti ili u većem postotku rezervacije koja nije realizirana nije opcija hotelskom menadžmentu. Jedan od pristupa kojim bi se zadržala fleksibilnost prema gostima, a istodobno minimizirali gubitci je razvijanje modela predviđanja otkaza. Takvim pristupom hotel može ponuditi na tržištu više smještajnih kapaciteta od stvarno dostupnih uz točnost do određenog postotka da će dio rezervacija biti otkazan, bolji prediktivni model znači posljedično manji reputacijski rizik kojem se hotel izlože zbog prekomjernog rezerviranja (overbooking-a). U nastavku rada razrađena je tema primjene strojnog učenja na problematiku otkaza rezervacija, u dijelu 2 opisani su podaci i rezultati eksploratorne analize promatranog skupa podataka, primijenjene metode i pristup rješavanju problema opisani su u dijelu 3, rezultati su opisani u dijelu 4 uz osvrt na druge pristupe rješavanju istog problema u dijelu 5. U dijelu 6 opisane su ideje mogućeg nastavka istraživanja.

II. OPIS PROBLEMA

A. Skup podataka

Skup podataka preuzet je sa stranice Kaggle pod nazivom Hotel booking demand dataset [1]. Sadrži podatke o rezervacijama dva hotela u Portugalu, hotel H1 je resort hotel u pokrajini Algarve, a H2 je gradski hotel u Lisabonu. Podaci za oba hotela imaju istu strukturu, 31 varijabla opisuje 40.600,00 zapisa hotela H1 i 79.330,00 zapisa hotela H2. Svaki zapis predstavlja rezervaciju u razdoblju od 01. srpnja 2015. do 31. kolovoza 2017. godine uključujući otkazane i realizirane rezervacije.

B. Analiza I priprema podataka

Analizom odabranog skupa promatrani su podaci, značajke i uočena je potreba za transformacijom podataka.

Atributi:

ADR- prosječna dnevna zarada po rezervaciji

Adults- broj odraslih

Agent – ID putničke agencije koja je napravila rezervacija

ArrivalDateDayOfMonth – dan datuma dolaska

ArrivalDateMonth – mjesec datuma dolaska

ArrivalDateWeekNumber –

ArrivalDateYear – godina datuma dolaska

AssignedRoomType – tip sobe

Babies – broj beba

BookingChanges – broj izmjena u rezervaciji

Children – broj djece

Company – ID firme koja je „gost“

Country – država gosta

CustomerType – ugovor, grupa, pojedinačna povezana s drugom rezervacijom, pojedinačna nepovezana s drugom rezervacijom

DaysInWaitingList – ukupno dana na listi čekanja prije uspješne realizacija

DepositType – bez depozita, nepovratni depozit (puna cijena), povratni depozit (dio cijene)

DistributionChannel – TA (Putnička agencija) ili TO (Operatori putovanja)

IsCanceled – otkazana ili realizirana rezervacija

IsRepeatedGuest – gost koji je već imao rezervacije

LeadTime – broj dana između datuma unosa rezervacija i datuma rezervacije

MarketSegment – Travel agency ili Tour operator

Meal – obrok koji je uključen u rezervaciju: bez obroka, doručak, doručak i još jedan obrok, tri obroka

PreviousBookingsNotCanceled - broj prethodno realiziranih rezervacija gosta

Identify applicable funding agency here. If none, delete this text box.

PreviousCancellations –broj prethodno otkazanih rezervacija gosta

RequiredCardParkingSpaces – broj parkirnih mjesta zahtijevanih od gosta

ReservationStatus – zadnji status rezervacije: otkazano, rezelizirano, No show (gost se nije pojavio)

ReservationStatusDate – datum zadnje promjene zapis rezervacije (odjava gosta, nije se pojavio, datum otkaza)

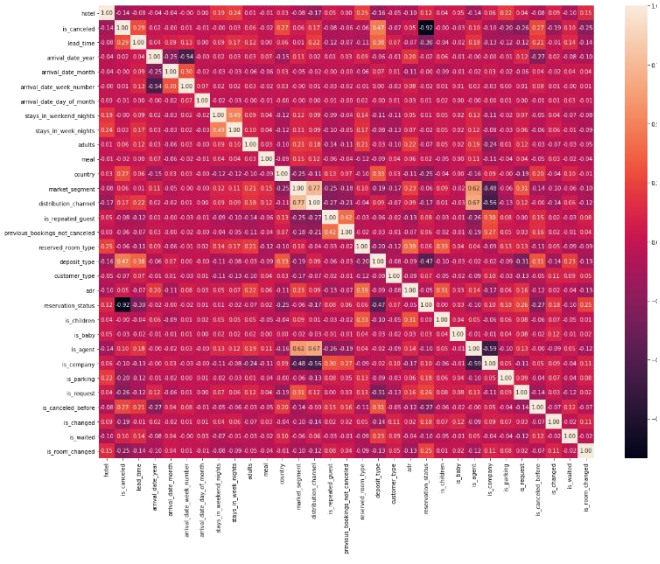
ReservedRoomType – kod tipa sobe

StaysInWeekendNights – broj vikenda uključenih u rezervaciju

StaysInWeekNights – broj dana u tjednu uključenih u rezervaciju

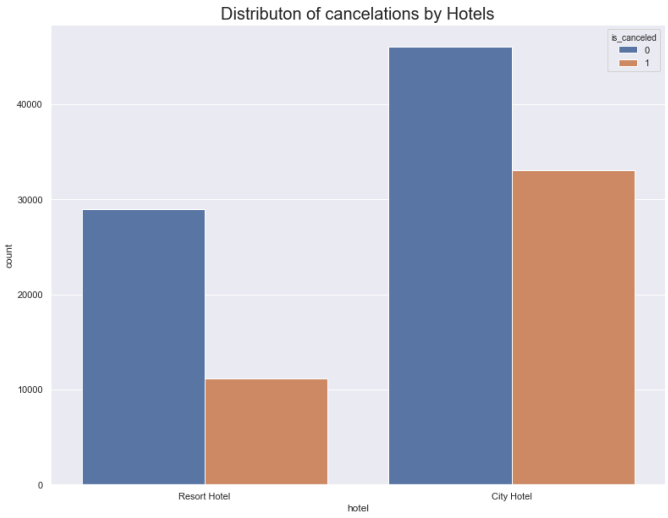
TotalOfSpecialRequests – broj posebnih zahtjeva

U dijelu pripreme iz skupa podataka izbačeni su zapisi u kojima je broj osoba koje su napravile rezervaciju bio nula. Za atribute *agent, company i children* null vrijednosti zamijenjene su s cijelim brojem 0, a za atribut *country* tamo gdje nema vrijednosti, postavljeno je mode najčešćom. Stvorene su nove binarne varijable kako bi se lakše analizirali odnosi, koje samo prate postojanje vrijednosti većih od nula kod aributa *children, babies, agent, company, required_car_parking_spaces, previous_cancellations, booking_changes, days_in_waiting_list*, te varijabla *is_room_cnaged* koja prati je li rezervirana soba drugačija od dodijeljene. Uklonjeni su atributi reservation *status i* reservation_*status_date* za koje je uočena visoka *koreliranost* sa varijablom *is_canceled* (koreliranost >90%), te *market_segment* zbog korelacije sa *distribution_channel*.



Slika 1 Matrica korelacije prije uklanjanja varijabli

Analizom podataka uočeno je da gradski hotel ima ukupno više rezervacija, ali i veći postotak otkazanih rezervacija. Oba tipa hotela popularnija su u ljetnim mjesecima, ali resort hotel ima manji pad broja gostiju tokom zimskih mjeseci. U analizi odnosa ostalih varijabli s obzirom na promatranu varijablu is_*canceled*, otkazivanja rezervacija učestalija su u ljetnim mjesecima za resort hotel, dok za gradski hotel nema značajnih razlika u broju otkazanih rezervacija po mjesecima. Zimski mjeseci imaju nešto niži postotak otkazivanja, ali je to za gradske hotele zanemarivo, dok je za resorte nešto izraženije. Zaključujemo da gradski hoteli imaju podjednak postotak otkazanih rezervacija tijekom cijele godine, dok resorti imaju više otkazivanja u ljetnim mjesecima. Očekivano za oba hotela, najčešće se otkazuju online rezervacije, a za oba tipa hotela otkazivanje je vjerojatnije što je više vremena preostalo do rezervacije. Prosječno vrijeme otkazivanja je nešto više od 100 dana do rezervacije. Nema značajne razlike kod otkazivanja rezervacije obzirom imaju li gosti djecu, ali gosti s bebama rjeđe otkazuju. Gosti koji koriste parking u pravilu ne otkazuju rezervaciju, slično kao i gosti koji imaju posebne zahtjeve te oni koji mijenjaju rezervaciju. Gosti koji koriste agenta otkazuju rezervaciju češće nego ostali, ali najveća razlika je kod gostiju koji su prethodno otkazivali rezervaciju, kojih čak više od 80% otkazuje rezervaciju.



Slika 2 Distribucija otkazanih rezervacija po hotelima

III. METODE I PRISTUP RJEŠAVANJU PROBLEMA

U rješavanju klasifikacijskog problema promatrani su rezultati četiri algoritma opisana u daljnjem tekstu.

A. Decision tree

Metoda slučajnih šuma (eng. random forest) generira ansambl stabla odlučivanja (engl. decision trees) i uprosječuje njihova predviđanja. Svako stablo je izgrađeno na slučajnom podskupu (s ponavljanjem) skupa za učenje i prilikom izgradnje svakog stabla se uvijek uzima slučajni podskup značajki za izgradnju svakog pojedinog čvora. Paranetar n_estimators regulira broj stabala dok parametar max_features regulira broj nasumično odabranih značajki koje se razmatraju prilikom izgradnje svakog čvora. Parametar max_depth regulira maksimalnu dubinu stabala u ansamblu.

B. *Random forest*

Random Forest algoritam za multi-class klasifikaciju temeljen na stablima odluke. Stablo odluke klasificira objekt pitajući da-ne , tj. 0 - 1 pitanja. Dobro formirano stablo svakim će pitanjem prepoloviti broj opcija, čime se i za veliki broj opcija vrlo brzo dolazi do odluke. Međutim, kod stabala odlučivanja već i na malim dubinama (npr. 5) dolazi do problema overfitting -a i dva stabla odlučivanja mogu davati međusobno vrlo nekonzistentne odluke. Pokazuje se da će davati konzistentne odluke na onim područjima u podacima koja nisu obuhvaćena overfittingom, a razlikovat će se u onima koja jesu. Dakle, ukoliko imamo informacije od više stabala s problemom overfitting- a, kombiniranjem njihovih odluka možemo zapravo riješiti taj problem, što i čini Random Forest metoda. Metoda Random Forest primjenjuje se na odabranim podacima, prvo se trenira model na skupu za treniranje, a zatim radi predikcija na testnom skupu.

C. *Logistic regression*

Logistička regresija odnosno logistički model se koristi za predviđanje vjerojatnosti događaja pomoću prilagođavanja podataka logističkoj krivulji prepoznatljivoj po svojemu S-obliku. Poveznica kojom je određen model logističke regresije je logit funkcija kojom se djeluje na vjerojatnost uspjeha u binomnoj distribuciji. Logistička regresija predstavlja vrstu regresijske analize u kojoj je zavisna (odzivna) varijabla dihotomna, odnosno binarna i kodira se s 0 ili 1 te postoji najmanje jedna nezavisna odnosno prediktorska varijabla. Navedeno u stvarnosti predstavlja modeliranje bilo kojeg problema kod kojeg se ciljni događaj može prevesti u kategorijsku varijablu (da/ne).

D. *XGBoost*

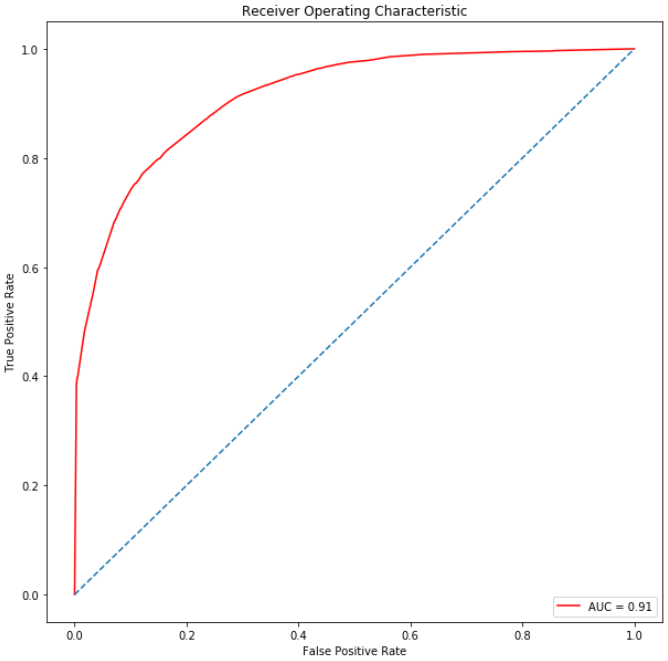
XGBoost (puni naziv Extreme Gradient Boosting) predstavlja brz i efikasan model za rad s tabličnim podacima. Koristi se za nadzirano učenje i multi-class klasifikaciju. XGBoost je implementacija Gradient Boosted Decision Trees algoritma, sastoji se od ciklusa koji grade nove modele i kombiniraju ih u već postojeći model. Potrebne su osnovne predikcije za pokretanje ciklusa pa je polazna točka naivan model koji je dan na ulazu. U praksi se pokazalo da početno predviđanje ne mora biti potpuno točno, jer će naknadne nadopune modela rješavati pogreške. Ciklus započinje računanjem pogrešaka za svako promatranje u skupu podataka. Zatim se izrađuje novi model koji će predviđati te pogreške. Taj model se zatim dodaje u skup dosadašnjih modela. Da bismo napravili predikciju, potrebno je koristiti predikcije iz svih do sad formiranih modela. Pomoću tih predikcija, može se izračunati nove pogreške, izraditi sljedeći model i dodati ga prethodnim modelima.

IV. REZULTATI

Prikazati ćemo sve dobivene rezultate te njihovu usporedbu.

A. *Decision tree*

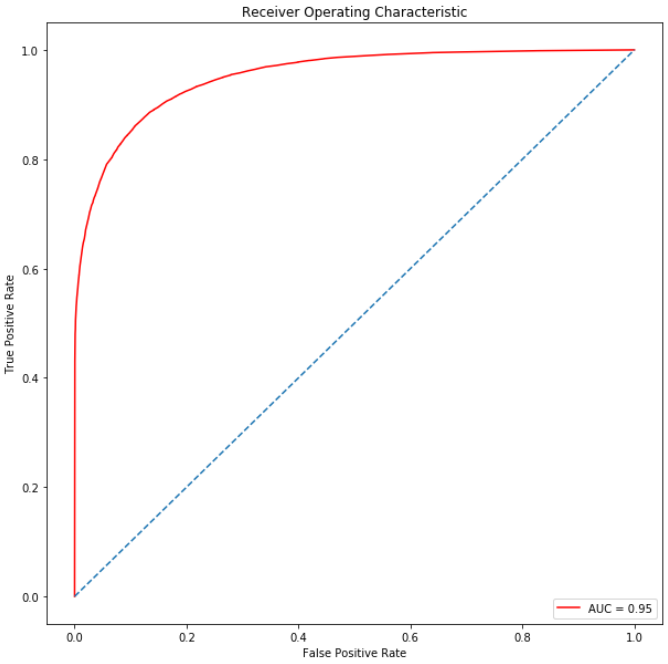
Korištenjem metode Decision tree dobili smo točnost od 84.12% uz AUC ROC od 82.15%.



Slika 3 Decision tree ROC krivulja

B. *Random forest*

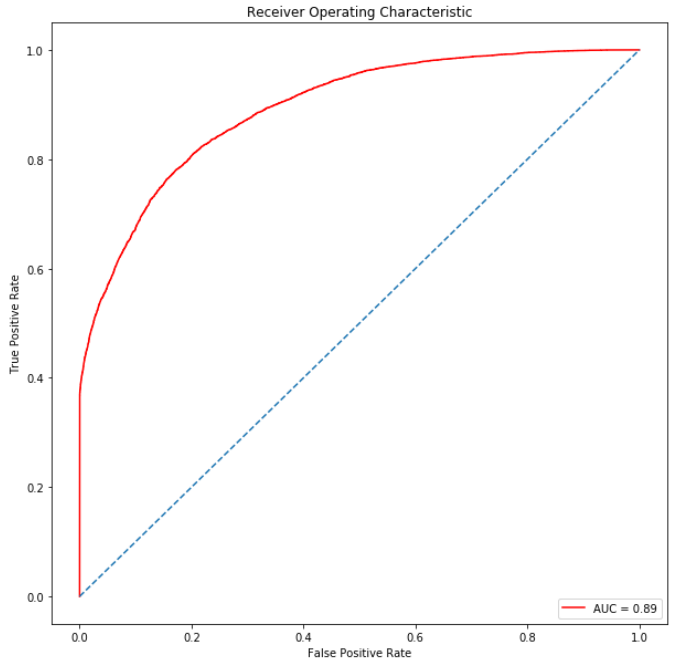
Korištenjem metode Random forest dobili smo točnost od 88.58% uz AUC ROC od 86.73%.



Slika 4 Random forest ROC krivulja

C. *Logistic regression*

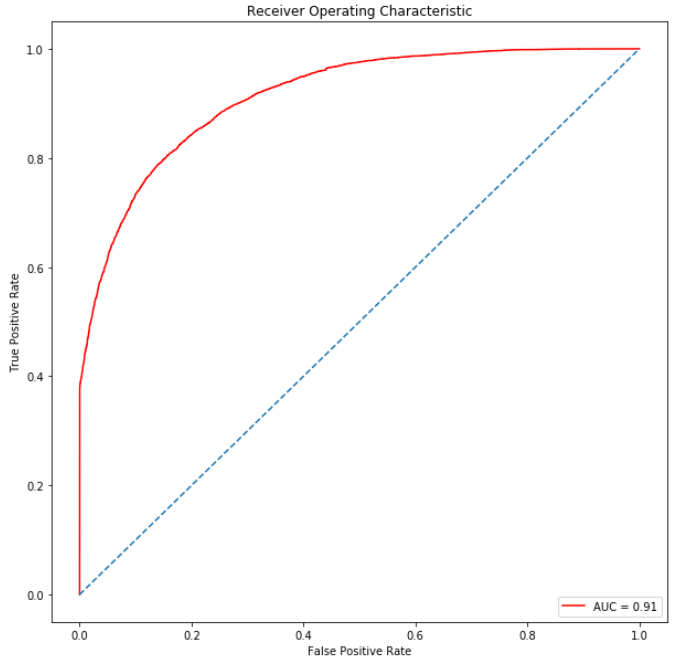
Korištenjem metode Logistic regression dobili smo točnost od 81.51% uz AUC ROC od 78.15%.



Slika 5 Logistic regression ROC krivulja

D. XGBoost

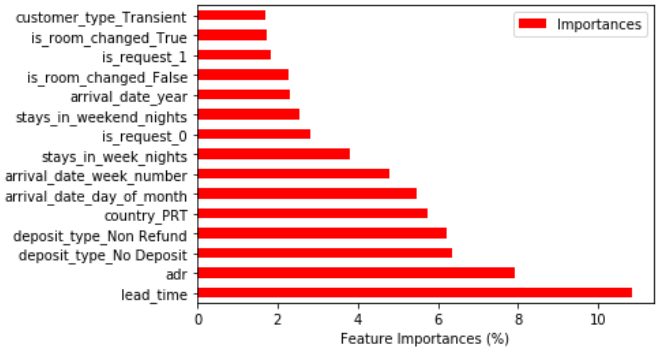
Korištenjem metode XGBoost dobili smo točnost od 81.51% uz AUC ROC od 78.15%.



Slika 6 XGBoost ROC krivulja

Zaključak nakon svih rezultata:

Vidimo iz priloženih rezultata da se najboljim pokazala metoda Random forest. Također smo provjerili hoće li se i za koliko promijeniti točnost ako podatke razdvojimo po hotelima uz Random forest metodu. Dobili smo rezultate unutar 2% od spojenih, što sugerira da bi model mogao biti dovoljno općenit i za primjenu na drugim hotelima. Provjerili smo još i važnost featurea kod Random foresta te smo dobili da je lead_time, što označava vrijeme između datuma rezervacije i datuma dolaska najznačajniji feature. Značajnost ostalih featurea se može vidjeti u priloženoj slici.



Slika 7 Random forest feature importance

V. OSVRT NA DRUGE PRISTUPE

U ovom radu primijenjen je drugačiji pristup s obzirom na ostale radove gdje su algoritmi strojnog učenja primijenjeni na svakom hotelu zasebno. Samim tim rezultati su bili precizniji, ali cilj ovog projekta je bio primijeniti modele na skupu podataka koju obuhvaća zapise oba hotela, kako bi se provjerilo kolika preciznost se može dobiti generaliziranim pristupom.

U projektu [2] na skupu podataka četiri resort hotela portugalske pokrajine Algarve primijenjene su metode Boosted Decision Tree, Decision Forest, Decision Jungle, Locally Deep Support Vector Machine i Neural Network, k-fold cross-validacija je korištena za testiranje performansi modela. Modeli su primijenjeni za svaki hotel zasebno, a Decision forest postiže najbolje rezultate u smislu preciznosti na tri od četiri hotela.

Studentski projekt [3] ima sličan pristup kao projekt [2], analizira se skup od gotovo 1,3M zapisa za 7 različitih hotela. Koriste se modeli: Naive Bayes, Logistic Regression, Decision tree i Random forest. Performanse modela provjeravaju se k-fold cross validacijom. Zbog razlike veće od 15% u postotku otkazanih rezervacija među hotelima i u ovom pristupu, modeli se primjenjuju za svaki hotel posebno. Random forest daje najbolje rezultate, a odmah poslije njega su rezultati postignuti Decision tree modelom.

VI. MOGUĆI NASTAVAK ISTRAŽIVANJA

Ideje za nastavak i poboljšanja:

- Primjena generaliziranog pristupa jednog prediktivnog modela na podacima više hotela koji koriste različite skupove podataka pa samim tim imaju različite zapise (ne jednako detaljne)
- Primjena modela na više različitih hotela, ali istog tipa, primjerice gradski hoteli koji se nalaze u različitim gradovima.
- Rezultat primjene modela na većem skupu podataka imao (više zapisa za hotel H1 i H2).

- Primjena modela na novijim podacima istih hotela (primjerice razbolje 2018.-2020. godine)
- Daljnja istraživanja mogla bi koristiti dodatne atribute, poput informacija o vremenskim prilikama (utjecaj najavljenih vremenskih prilika/neprilika na postotak otkaza rezervacija u oba tipa hotela) i osobnim podatcima gostiju (dob, spol, zanimanje, građanski status, broj djece, godišnji prihod).

a.

.

REFERENCES

[1] Dataset: https://www.kaggle.com/jessemostipak/hotel-booking-demand

[2] https://www.researchgate.net/publication/310504011_Predicting_Hotel_Booking_Cancellation_to_Decrease_Uncertainty_and_Increase_Revenue

[3] https://beta.vu.nl/nl/Images/werkstuk-leeuwen_rik_van_tcm235-876479.pdf

[4] https://www.sciencedirect.com/science/article/pii/S2352340918315191#bib5

[5] https://web.math.pmf.unizg.hr/nastava/su/materijali/

[6] https://towardsdatascience.com/predicting-hotel-cancellations-with-machine-learning-fa669f93e794