Hotel Booking Cancellation Prediction

Mitić Andrea, IN 30/2020, mitic.in30.2020@uns.ac.rs
Pantić Nikola, IN 40/2020, pantic.in40.2020@uns.ac.rs

I. Uvod

Hotelska industrija se suočava s rastućim izazovima u upravljanju rezervacijama. S obzirom na potencijalni uticaj otkazivanja na finansijsku stabilnost i planiranje resursa, razumevanje faktora i verovatnoće otkazivanja postaje ključno za hotelijere radi unaprednog planiranja strategija.

Kroz analizu podataka i primenu modela mašinskog učenja, hoteli mogu dublje razumeti obrasce ponašanja gostiju i efikasnije upravljati rezervacijama. Ovaj rad istražuje kako primena odgovarajućih modela mašinskog učenja može unaprediti predviđanje otkazivanja rezervacija, omogućavajući hotelima da pravovremeno reaguju i minimiziraju gubitke.

II. Baza podataka

Ova naučna studija koristi bazu podataka koja pruža uvid u potencijalne faktore koji mogu dovesti do otkazivanja hotelskih rezervacija. U bazi se nalazi okvirno 36 hiljada uzoraka sa 17 obeležja od kojih je jedno numeričko a preostala su kategorička. Samim tim ćemo se baviti klasifikacijom obeležja. Neka od obeležja su broj gostiju, podeljeni u broj odraslih i broj dece, broj noćenja u toku radnih dana i vikenda, prosečna cena rezervacije, status rezervacije itd.

III. Analiza i modeli

Od 17 obeležja koje baza podataka poseduje, dva smatramo redudantnim:

1. BookingID - ID rezervacije
2. P-not-C - da li rezervacija nije bila otkazana (već posedujemo informaciju da li je otkazana)

Koristićemo sledeća tri modela kako bismo rešili problem klasifikacije:

1. KNN (K-Nearest Neighbors)
2. SVM (Support Vector Machines)
3. LDA (Linear Discriminant Analysis)

Da bismo odabrali najbolje hiperparametre za svaki model, koristićemo GridSearchCV na trening i validacionim skupovima pre nego što poredimo test skupove krajnjim rešenjima.

K-najbližih suseda (KNN) je jednostavan, ali efikasan algoritam za klasifikaciju i regresiju. Osnovna ideja KNN-a je da se novi podaci klasifikuju na osnovu sličnosti sa najbližim podacima u trening skupu. Ovaj algoritam ne zahteva pretpostavke o distribuciji podataka i može biti veoma koristan za probleme sa linearno nerazdvojivim podacima.

Za K-najbližih suseda (KNN) algoritam odabrali smo parametre koje smo ispitivali u GridSearchCV zbog njihove ključne uloge u performansama modela. Broj suseda (n_neighbors) utiče na kompleksnost modela - manji broj suseda može dovesti do preprilagođavanja, dok veći broj suseda može dovesti do potprilagođenja. Metrika (metric) određuje meru udaljenosti između tačaka u prostoru atributa. Odabrali smo 'hamming', 'euclidean' i 'manhattan' jer svaka od njih može bolje odgovarati različitim vrstama podataka i različitim topologijama prostora.

Mašine na bazi vektora (SVM) su moćan algoritam za klasifikaciju i regresiju koji rade tako što pronalaze optimalnu hiperravan koja razdvaja podatke različitih klasa u prostoru sa što većom marginom. SVM je posebno efikasan u visokodimenzionalnim prostorima i može se prilagoditi različitim tipovima kernela radi efikasnijeg razdvajanja podataka.

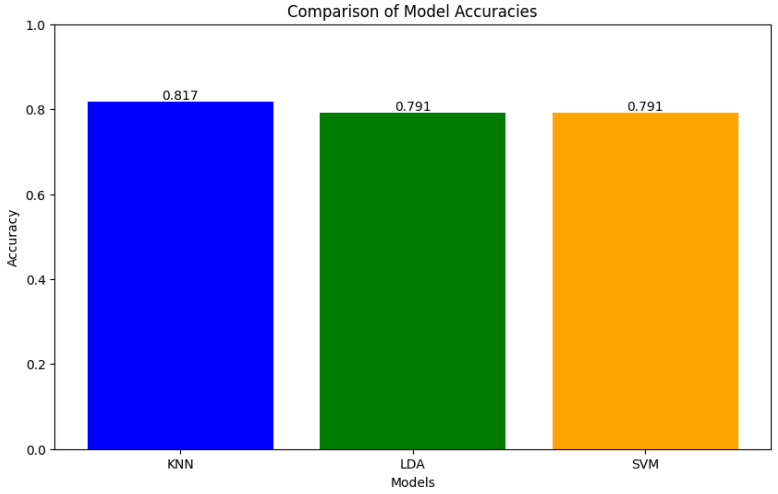
Za Support Vector Machine (SVM) algoritam, parametar konstante regularizacije (C) je ključan jer kontroliše težinu kazne za pogrešno klasifikovane tačke. Odabrali smo vrednosti [0.1, 1, 10] kako bismo istražili kako različite vrednosti C utiču na performanse modela. Kernel funkcija (kernel) je takođe bitna jer određuje oblik odlučujuće granice između klasa. 'linear' kernel se koristi kada su podaci linearno separabilni, 'rbf' kernel je višedimenzionalni i može se prilagoditi kompleksnijim oblicima podataka, dok 'poly' kernel koristi polinomijalnu funkciju za mapiranje podataka u višedimenzionalni prostor. Parametar gama (gamma) utiče na širinu odlučujuće granice i odabrali smo 'scale' kako bi se automatski skalirala vrednost gama.

Linearna diskriminativna analiza (LDA) je statistička tehnika koja se koristi za pronalaženje linearnih kombinacija obeležja koje najbolje razdvajaju različite klase u podacima. LDA pokušava da modelira razlike između klasa u skupu podataka maksimizirajući razliku između srednjih vrednosti klase i minimizirajući varijabilnost unutar svake klase.

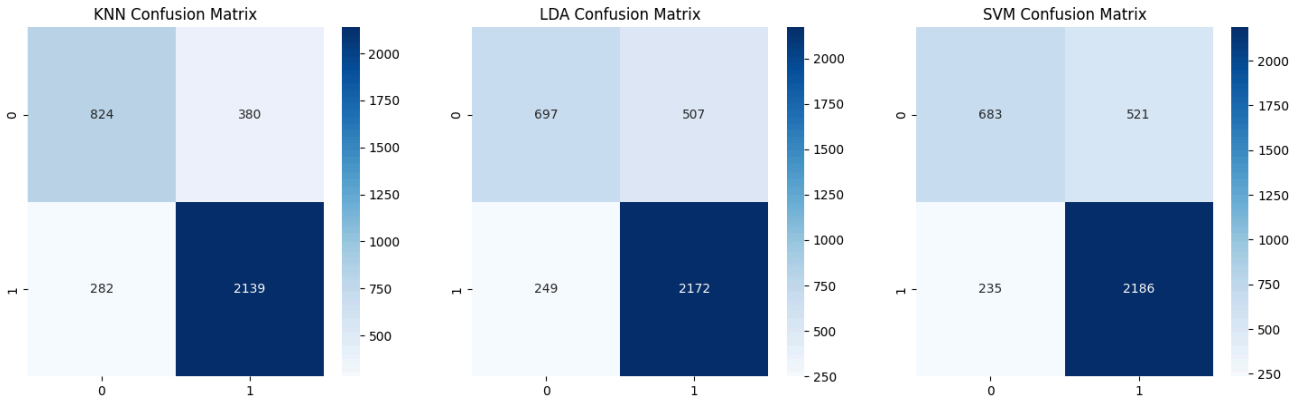
Za Linearnu diskriminativnu analizu (LDA), odabrali smo parametre koji se odnose na metod rešavanja (solver), smanjenje (shrinkage) i broj komponenata (n_components). 'lsqr' je odabran za solver jer je efikasan za rešavanje problema sa visokom dimenzionalnošću. Parametar shrinkage razmatran je kako bismo videli da li primena smanjenja može poboljšati performanse modela, dok je parametar n_components istraživan kako bi se testirala različita dimenzionalnost prostora atributa.

IV. Rezulatati

Od primenjenih metoda, KNN se pokazao kao najefikasniji i najpreciznji model za klasifikaciju naše baze podataka, sa estimacijom od 81.99%, dok su SVM i LDA ostvarili identične procene od 79.87%.



Slika 1: Poređenje preciznosti modela



Slika 2: Matrice konfuzije modela

V. Zaključak

Naša studija pruža dragocen uvid u primenu mašinskog učenja u hotelskoj industriji i ističe važnost analize podataka u unapređenju poslovnih procesa i donošenju informisanih odluka. Dalji rad u ovom domenu može doprineti razvoju efikasnijih alata i tehnika za upravljanje rezervacijama i optimizaciju poslovanja u hotelskoj industriji.