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 kako bi unapred planirali strategije.

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, koji su 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. Stabla odluke (Decision Trees)

Takođe, prvo ćemo izvršiti poređenja rešenja sa i bez redukcije dimenzionalnosti korišćenjem 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.

Stabla odluke su algoritmi mašinskog učenja koji koriste logičke testove na atributima podataka kako bi klasifikovali ili predvideli rezultate. Oni organizuju podatke u stablo sa čvorovima koji predstavljaju testove atributa i listovima koji sadrže klasifikacije ili predikcije.

Za Decision Trees klasifikator, korišćenjem GridSearchCV-a, istraživali smo različite kombinacije parametara kako bismo pronašli optimalni model stabla odluke. U ovom kontekstu, pažljivo smo razmatrali kriterijume 'gini' i 'entropy', koji predstavljaju mere nečistoće podataka i utiču na način kako stablo bira svoje odluke. Gini kriterijum se fokusira na verovatnoću pogrešne klasifikacije ako se slučajno izabere tačka iz skupa podataka, dok entropija se računa kao suma verovatnoća pojavljivanja svake klase pomnožena sa logaritmom te verovatnoće. Takođe, istraživali smo dubinu stabla, što određuje koliko daleko stablo može ići u svojoj strukturi, kao i minimalne brojeve uzoraka potrebne za listove čvorova kao i minimalne brojeve uzoraka potrebne za podelu čvorova.

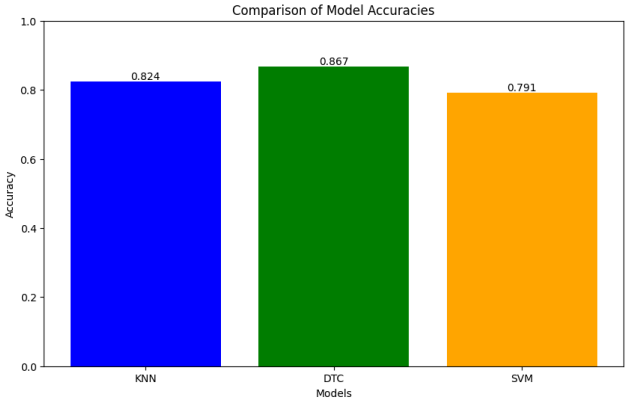
IV. Rezultati

1. Rezultati bez LDA:

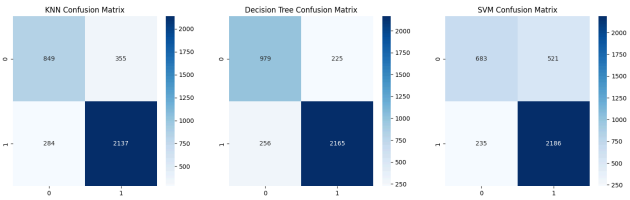
Korišćenjem GridSearchCV, izabrali smo sledeće parametre za finalne modele:

- Za KNN model:
  - Broj suseda (n_neighbors): 11
  - Metrika (metric): ‘hamming’
- Za SVM model:
  - Regularizacioni parametar (C): 0.1
  - Kernel: ‘linear’
  - Gama parametar: ‘scale’
- Za model stabla odluke:
  - Maksimalna dubina stabla (max_depth): 10
  - Kriterijum podela (criterion): ‘gini’
  - Minimalni broj uzoraka za list (min_samples_leaf): 1
  - Minimalni broj uzoraka za podelu (min_samples_split): 2

Na osnovu dobijenih rezulata, primećujemo da je klasifikator stabla odluke imao najbolju performansu sa 86.7% tačnosti, dok je k-najbližih suseda zauzeo drugo mesto sa 82.4% tačnosti, a mašine na bazi vektora nosača, čak iako najgori model, dokazao se kao pouzdan klasifikator sa 79.1% tačnosti.



Slika 1.1: Poređenje preciznosti modela bez LDA



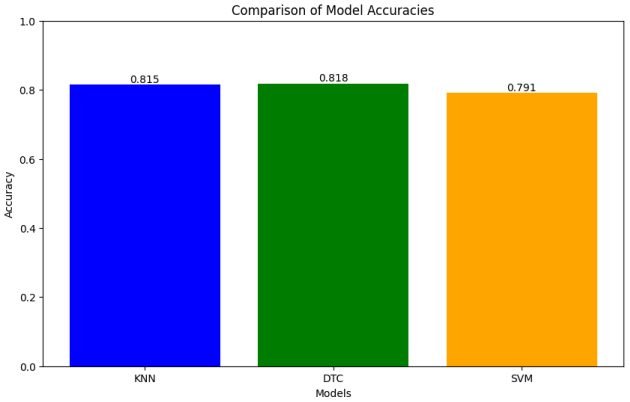
Slika 1.2: Matrice konfuzije modela bez LDA

2. Rezultati sa LDA:

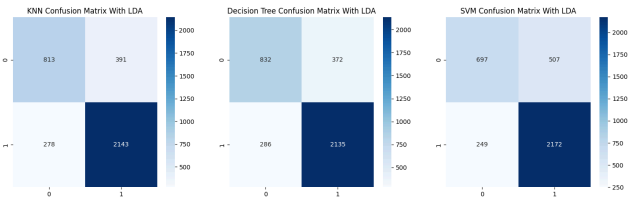
Korišćenjem GridSearchCV, izabrali smo sledeće parametre za finalne modele:

- Za KNN model:
  - Broj suseda (n_neighbors): 7
  - Metrika (metric): ‘euclidean’
- Za SVM model:
  - Regularizacioni parametar (C): 10
  - Kernel: ‘rbf’
  - Gama parametar: ‘scale’
- Za model stabla odluke:
  - Maksimalna dubina stabla (max_depth): 20
  - Kriterijum podela (criterion): ‘entropy’
  - Minimalni broj uzoraka za list (min_samples_leaf): 1
  - Minimalni broj uzoraka za podelu (min_samples_split): 2

Na osnovu dobijenih rezulata, primećujemo da je ponovo klasifikator stabla odluke imao najbolju performansu sa 81.8% tačnosti, dok je model k-najbližih suseda zauzeo drugo mesto sa 81.5% tačnosti, a model mašina na bazi vektora nosača, čak iako najgori model, dokazao se kao pouzdan klasifikator sa 79.1% tačnosti.



Slika 2.1: Poređenje preciznosti modela sa LDA



Slika 2.2: Matrice konfuzije modela sa LDA

Analizirajući naše rezultate, primetili smo da trenutno nema koristi od smanjenja dimenzionalnosti podataka pomoću LDA, modeli daju lošije procene od procena sa originalnim obeležjima. Ovo ističe potrebu za daljim istraživanjem i prilagođavanjem metodologije kako bismo bolje razumeli karakteristike naših podataka i primenili odgovarajuće tehnike mašinskog učenja.

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 efikasnih alata i tehnika za upravljanje rezervacijama i optimizaciju poslovanja u hotelskoj industriji.