Primjena strojnog učenja u predviđanju cijena avionskih letova

Lopotar, Matej

Master's thesis / Diplomski rad

2024

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Electrical Engineering and Computing / Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:168:777466

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2025-04-01**

Repository / Repozitorij:

FER Repository - University of Zagreb Faculty of Electrical Engineering and Computing repozitory

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 368

PRIMJENA STROJNOG UČENJA U PREDVIĐANJU CIJENA AVIONSKIH LETOVA

Matej Lopotar

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 368

PRIMJENA STROJNOG UČENJA U PREDVIĐANJU CIJENA AVIONSKIH LETOVA

Matej Lopotar

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

DIPLOMSKI ZADATAK br. 368

Pristupnik:	**Matej Lopotar (0036523737)**
Studij:	Računarstvo
Profil:	Znanost o podacima
Mentor:	izv. prof. dr. sc. Tomislav Burić

Zadatak: **Primjena strojnog učenja u predviđanju cijena avionskih letova**

Opis zadatka:

U današnje vrijeme aviokompanije koriste složene strategije i algoritme za dinamičko određivanje cijena letova ovisno o raznim financijskim, marketinškim i ljudskim faktorima. Tema ovog rada je predviđanje kretanja cijena avionskih letova koristeći metode strojnog učenja. Potrebno je proučiti i implementirati nekoliko različitih predikcijskih modela koji se temelje na regresijskim stablima odlučivanja. Dobivene rezultate je potrebno detaljno analizirati i prikazati grafički te napraviti usporedbu točnosti korištenih modela.

Rok za predaju rada: 28. lipnja 2024.

Sadržaj

1. Uvod

Povijest zračnog prometa počinje s prvim uspješnim letom braće Wright 1903. godine, što je označilo početak moderne avijacije. Tijekom desetljeća, zračni promet se razvijao iz eksperimentalne faze u globalnu industriju omogućujući brza i učinkovita putovanja preko cijelog svijeta.

U današnjem svijetu zračni promet najbrži je način putovanja te igra ključnu ulogu u povezivanju ljudi, njihovih kultura i ekonomija. Ljudi su konačno opet slobodni putovati po svijetu bez ikakvih ograničenja koja su postojala za vrijeme pandemije koronavirusa. Aviokompanije koriste složene strategije i algoritme za dinamičko određivanje cijena letova ovisno o raznim financijskim, marketinškim i ljudskim faktorima. Iz tog razloga cijene aviokarata mogu se drastično mijenjati u kratkim vremenskim razdobljima zbog promjena u potražnji, konkurenciji, troškovima goriva i slično. Predvidivi parametri koji se koriste u dinamičkom određivanju cijena uključuju broj dana do leta, dan leta, praznike, potražnju na ruti, raspoloživost sjedala, udaljenost leta, klasu putnika i povijesne podatke o cijenama. Osim toga, većinom nepredvidivi parametri kao što su vremenski uvjeti, događaji u odredišnim gradovima, pojava neočekivanih događaja ili politička situacija u zemlji također utječu na cijene karata. [1] Ljudi s druge strane traže najpovoljnije cijene karata pa tako predviđanje cijena avionskih karata postaje sve važnija kod donošenja odluka o tome treba li kupiti kartu odmah ili čekati nižu cijenu. Pretpostavlja se da je bolje uvijek kupiti kartu puno ranije što ne mora biti istina - aviokompanije mogu smanjiti cijene karata kasnije kako bi povećale prodaju i na primjer popunile avion.

Strojno učenje je jedna od najpopularnijih istraživačkih tema u računalnim znanostima i inženjerstvu, koja je primjenjiva u mnogim disciplinama. [2] Algoritmi strojnog učenja koji se temelje na regresijskim stablima odlučivanja korišteni su u predviđanju cijena. U ovom radu istražena je primjena metoda strojnog učenja za predviđanje cijena

avionskih karata.

Razumijevanje i predviđanje cijena avionskih karata ne koristi samo putnicima, nego i aviokompanijama koje mogu povećati popunjenost avionskih letova, a samim time i prihode. Precizni modeli za predviđanje cijena mogu pomoći putnicima u odabiru optimalnog vremena za kupovinu karata, smanjujući troškove putovanja i povećavajući zadovoljstvo korisnika.

Cilj je implementirati i evaluirati nekoliko različitih predikcijskih modela temeljenih na regresijskim stablima odlučivanja kao što su *RandomForestRegressor, ExtraTreesRegressor i DecisionTreeRegressor*, detaljno analizirati dobivene rezultate te ih grafički prikazati, usporediti točnost korištenih modela te odrediti najbolji model.

2. Stabla odlučivanja

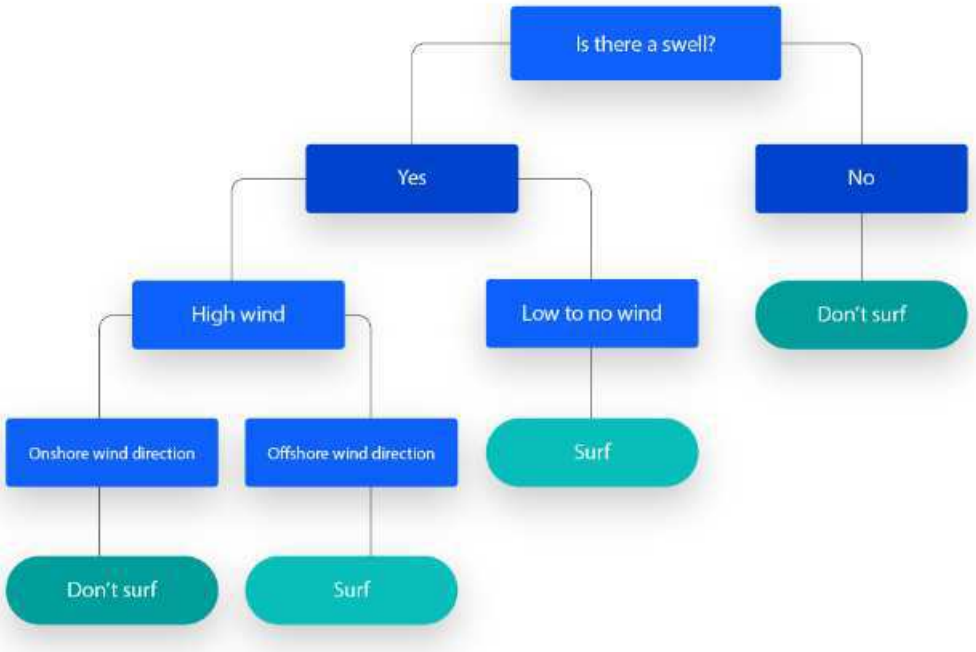
Stabla odlučivanja su neparametarska metoda nadziranog učenja koja se koristi za klasifikaciju i regresiju. Cilj je stvoriti model koji predviđa vrijednost ciljane varijable učenjem jednostavnih pravila odlučivanja izvedenih iz značajki podataka.

U istraživanju su korištena tri predikcijska modela koja se temelje na regresijskim stablima odlučivanja - *RandomForestRegressor, ExtraTreesRegressor i DecisionTreeRegressor*.

Glavne prednosti stabala odlučivanja su to što su vrlo jednostavna za razumijevanje i interpretaciju i što postoji mogućnost vizualizacije. Sve te stvari omogućuju lakše shvaćanje i objašnjenje dobivenih rezultata. Kompjuterska složenost im je logaritamska i ovisi o broju podataka korištenih za treniranje stabla što ih čini efikasnima za velike skupove podataka. Još jedna velika prednost stabala odlučivanja je mogućnost rada s dvije vrste podataka - kategorijskim podacima i numeričkim podacima. Klasifikacijsko stablo odlučivanja koristi se za kategorijske varijable dok se regresijsko stablo odlučivanja koristi za numeričke varijable. Za razliku od neuronskih mreža koje su svrstane u "black box" modele, stabla odlučivanja su "white box" modeli što znači da su transparentni i da omogućuju korisnicima da shvate kako modeli funkcioniraju. [3] Iako imaju puno prednosti, stabla odlučivanja naravno imaju i neke nedostatke.

Jedan od nedostataka je stvaranje previše kompleksnih stabala koja ne generaliziraju dobro podatke. Mehanizmi poput obrezivanja (*engl. pruning*), postavljanja minimalnog broja uzoraka potrebnih na listu ili postavljanja maksimalne dubine stabla su neophodni kako bi se izbjegao ovaj problem. Također, stabla odlučivanja mogu biti nestabilna jer male varijacije u podacima mogu rezultirati potpuno drugačijim stablom. Ovaj problem se može ublažiti korištenjem stabala odlučivanja unutar ansambla. Problem

učenja optimalnog stabla odlučivanja je poznat kao NP-kompletan (*engl. nondeterministic polynomial-time complete*) pod nekoliko aspekata optimalnosti čak i za jednostavne koncepte. Kao posljedica toga, praktični algoritmi za učenje stabla odlučivanja temelje se na heurističkim algoritmima poput pohlepnog algoritma gdje se lokalno optimalne odluke donose na svakom čvoru. Takvi algoritmi ne mogu jamčiti povratak globalno optimalnog stabla odlučivanja. Ovo se može ublažiti treniranjem više stabala u ansambl učeniku, gdje se značajke i uzorci nasumično uzimaju s ponavljanjem. [3] Postoje koncepti koje je teško naučiti jer ih je nemoguće prikazati pomoću stabla odlučivanja kao što je XOR problem. Također, algoritmi za stablo odlučivanja kreiraju pristrana stabla ako neke klase dominiraju, stoga je preporučeno balansiranje skupa podataka prije treniranja stabla odlučivanja. [3]



Slika 2.1. Primjer stabla odlučivanja [4]

2.1. Klasifikacijsko stablo odlučivanja

Klasifikacijska stabla odlučivanja su jedna od najmoćnijih i najpopularnijih metoda u području rudarenja podataka i strojnog učenja. Ova tehnika koristi hijerarhijsku strukturu za donošenje odluka i predviđanje pripadnosti objekata određenim klasama. Stabla odlučivanja su jednostavna za razumijevanje i interpretaciju, što ih čini privlačnim alatom za istraživače i praktičare. [5]

Cilj klasifikacijskog stabla je objasniti ili predvidjeti odgovore na kategorijskoj zavisnoj varijabli. Klasifikacijsko stablo odlučivanja izgrađeno je od čvorova i bridova gdje svaki unutarnji čvor predstavlja test na nekoj prediktivnoj varijabli, svaki brid predstavlja ishod tog testa, a svaki list (terminalni čvor) predstavlja klasu ili distribuciju klase.

Proces izgradnje stabla započinje s cijelim skupom podataka, a zatim se postupno dijeli na podskupove na temelju vrijednosti prediktorskih varijabli. Podjele se vrše na način koji maksimizira razdvajanje između različitih klasa.

2.2. Regresijsko stablo odlučivanja

Regresijska stabla odlučivanja predstavljaju moćan i intuitivan alat u području strojnog učenja, posebno korisna za rješavanje regresijskih problema gdje je cilj predvidjeti kontinuirane vrijednosti. Ova metoda je proširenje koncepta stabala odlučivanja, koja se tradicionalno koriste za klasifikacijske zadatke

Svaka regresijska tehnika uključuje jednu izlaznu varijablu i jednu ili više ulaznih varijabli. Izlazne varijable su numeričke vrijednosti i u našem slučaju to je cijena avionskih karata. Standardni postupak izgradnje regresijskog stabla dopušta da ulazne varijable budu kombinacija kategorijskih i kontinuiranih varijabli. U radu su sve kategorijske varijable pretvorene u numeričke.

Stablo odluke se generira kada svaki čvor u stablu sadrži test na vrijednost određene ulazne varijable. Predviđene vrijednosti izlazne varijable nalaze se u terminalnim čvorovima stabla. Regresijsko stablo je dizajnirano da približi funkcije s realnim vrijednostima umjesto da se koristi za metode klasifikacije. Gradi se korištenjem procesa zvanog binarno rekurzivno dijeljenje. Ovo je iterativni proces koji dijeli podatke na particije ili grane, a zatim nastavlja dijeliti svaku particiju na manje skupine kako se metoda kreće prema gore kroz stablo po svakoj grani. Algoritam počinje raspoređivati podatke u prve dvije particije ili grane koristeći svaku binarnu podjelu na svakom polju.

Algoritam odabire podjelu koja minimizira zbroj kvadrata odstupanja od srednje vrijednosti u dvije odvojene particije. [5]

Neka su S_1 i S_2 dvije particije, tada je cilj minimizirati izraz:

$$\sum_{i \in S_1} (y_i - \bar{y}_{S_1})^2 + \sum_{i \in S_2} (y_i - \bar{y}_{S_2})^2 \tag{2.1}$$

gdje je $\bar{y}_{S_1}$ srednja vrijednost u S_1, a $\bar{y}_{S_2}$ srednja vrijednost u S_2.

Pravilo razdvajanja se zatim primjenjuje na svaku od novih grana. Proces se nastavlja dok svaki čvor ne dosegne minimalnu veličinu čvora koju je odredio korisnik i postane terminalni čvor. Čvor se također smatra terminalnim čvorom ako, iako nije dosegnuo minimalnu veličinu, zbroj kvadrata odstupanja od srednje vrijednosti u čvoru iznosi nula:

$$\sum_{i \in S} (y_i - \bar{y}_S)^2 = 0 \tag{2.2}$$

gdje je S čvor, a $\bar{y}_S$ srednja vrijednost čvora S. Ovo pravilo razdvajanja se zatim primjenjuje na svaku od novih grana. Proces se nastavlja dok svaki čvor ne dosegne minimalnu veličinu koju je odredio korisnik i postane terminalni čvor.

2.2.1. Regresor slučajnih šuma

Slučajne šume (*engl. Random Forests* učinkovit su alat u predikciji. Zakon velikih brojeva pokazuje da slučajne šume uvijek konvergiraju tako da pretreniranje modela nije moguće. Uvođenje odgovarajuće vrste nasumičnosti čini ih preciznim klasifikatorima i regresorima. Nadalje, okvir u smislu jačine pojedinačnih prediktora i njihovih korelacija daje uvid u sposobnost random šume za predikciju. Korištenjem out-of-bag procjene, konkretiziraju se inače teorijske vrijednosti jačine i korelacije. [6]

Regresor slučajnih šuma je napredni algoritam strojnog učenja koji se koristi za regresijske zadatke. Razvijen je kao proširenje osnovne ideje stabala odlučivanja kako bi poboljšao točnost i robusnosti modela. Slučajna šuma koristi zbirku stabala odlučivanja, odakle i potječe njegovo ime.

Algoritam se temelji na agregiranju predikcija više stabala odlučivanja. Svako pojedinačno stablo u šumi izrađeno je od različitih podskupova podataka koji se biraju slučajnim uzorkovanjem s ponavljanjem (bootstrap uzorkovanje). Ovo stvaranje višestrukih stabala rezultira modelom koji je manje sklon prenaučenosti i modelu koji generalizira

bolje skupu podataka za testiranje.

2.2.2. Regresor dodatnih stabala

Regresor dodatnih stabala (*engl. Extra Trees Regressor* poznatiji kao regresor ekstremno slučajnih stabala *engl.Extremely Randomized Trees Regressor* je varijanta stabla odlučivanja koja proširuje koncept nasumičnih šuma. Ova metoda koristi mnogo regresijskih stabala za izradu prediktivnog modela. Glavna ideja iza regresora ekstremno slučajnih stabala je dodatno povećati nasumičnost pri stvaranju stabala kako bi se smanjila varijanca modela i izbjegla prenaučenost pritom povećavajući točnost predikcije.

Algoritam dodatnih stabala (*engl. Extra Trees*, vrlo slično kao i algoritam slučajnih čuma, stvara mnoga stabla odlučivanja, ali je uzorkovanje za svako stablo nasumično, bez ponavljanja. Također, regresor dodatnig stabala koristi čitav originalni skup podataka za učenje umjesto da koristi podskupove kao što je to slučaj kod algoritma slučajnih šuma. Iz tog razloga pristranost modela je smanjena jer se koristi cijeli skup podataka, a ne neki podskup. Najvažnija karakteristika algoritma dodatnih stabala je nasumični odabir vrijednosti za dijeljenje značajke - umjesto da izračunava lokalno optimalnu vrijednost kao što je slučaj u algoritmu slučajnih šuma, dodatna stabla nasumično odabire vrijednost dijeljenja. To čini stabla raznolikima i nekoreliranim. Opisani algoritam ponavlja se više puta pa se tako stvori skup stabala. Predikcije svakog stabla se skupe kako bi se dobila konačna predikcija pomoću aritmetičke sredine u regresijskim problemima. Glavni cilj dodatnih stabala je smanjenje varijance i pristranosti. Isto tako proces izgradnje je brži jer nije potrebno za svaki čvor raditi procjene nego se čvor odabere nasumično. Tako je kompjuterska složenost stvaranja stabla $\mathcal{O}(n \log n)$ gdje je n veličina skupa za treniranje. [7]

Regresija dodatnih stabala je metoda koja koristi dodatnu nasumičnost za smanjenje varijance modela i poboljšanje točnosti predikcija. Ova metoda može biti korisna alternativa drugim metodama temeljenim na stablima, pružajući visoku točnost i računalnu učinkovitost u različitim kontekstima predikcije. Zbog svojih prednosti i jednostavnosti regresor dodatnih stabala postao je popularan alat u raznim područjima znanosti o podacima i strojnom učenju.

3. Analiza podataka

Za potrebe diplomskog rada korišteni su podaci tipa .csv koji su sadržavali sve podatke o avioprijevoznicima, letovima, cijenama i slično. Tablica ima sveukupno 62626 redaka i 10 stupaca kao što je prikazano na slici 3.1. Podaci sadrže njemačke letove u razdoblju od 25.10.2019. i 24.04.2020. godine. [8]

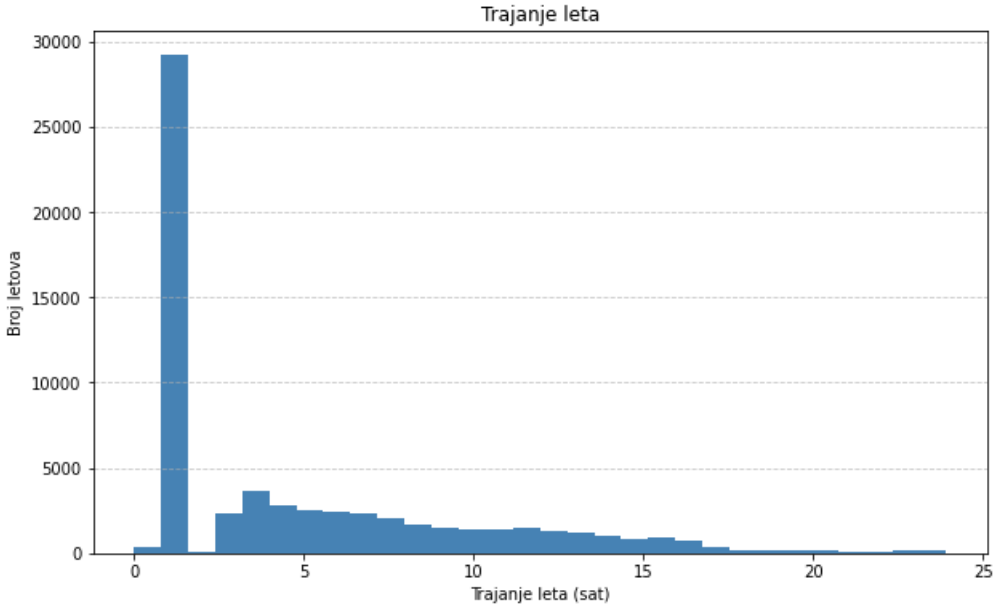
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 62626 entries, 0 to 62625
Data columns (total 10 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   departure_city          62626 non-null  object
 1   arrival_city            62626 non-null  object
 2   scrape_date             62626 non-null  object
 3   departure_date          62626 non-null  object
 4   departure_date_distance 62626 non-null  object
 5   departure_time          62626 non-null  object
 6   arrival_time            62626 non-null  object
 7   airline                 62626 non-null  object
 8   stops                   62626 non-null  object
 9   price (€)               62626 non-null  object
dtypes: object(10)
memory usage: 5.3+ MB
```

Slika 3.1. Prikaz svih stupaca, broja redaka i tipova podataka

Stupac *departure_city* predstavlja grad iz kojeg avion kreće dok stupac *arrival_city* predstavlja grad do kojeg avion leti. *scrape_date* je datum kupnje avionske karte dok je *departure_date* datum odlaska aviona, a *departure_date_distance* broj dana/tjedana/mjeseci između kupnje avionske karte i dana leta. *departure_time* je vrijeme polaska aviona, *arrival_time* je vrijeme dolaska aviona, *airline* je ime aviokompanije koja organizira let, *stops* je broj presjedanja i na kraju *price (€)* je cijena koju želimo predvidjeti. U sljedećim poglavljima bit će prikazane razne ovisnosti podataka, njihove raspodjele i statistika prije predviđanja cijena avionskih karata pomoću stabala odlučivanja.

3.1. Uređivanje podataka

U ovom poglavlju opisano je uređivanje podataka koje nisu u dobrom ili čitljivom formatu pa ih je potrebno promijeniti. Prvi problem pojavio se u stupcima *departure_time* i *arrival_time* u kojima su vremena bila zapisana u različitim formatima. Tako je dio podataka imao sufiks "am" ili "pm", a dio Uhr što na njemačkom znači "o'clock". Sva vremena promijenjena su u 24h format i zapisana u nove stupce. Nakon toga dodan je novi stupac u tablicu *flight_duration* koji predstavlja trajanje leta i izračunat je pomoću prethodno navedenih stupaca. Što se tiče "null" podataka oni u skupu nisu postojali pa nije bilo potrebno čistiti tablicu od tih podataka.

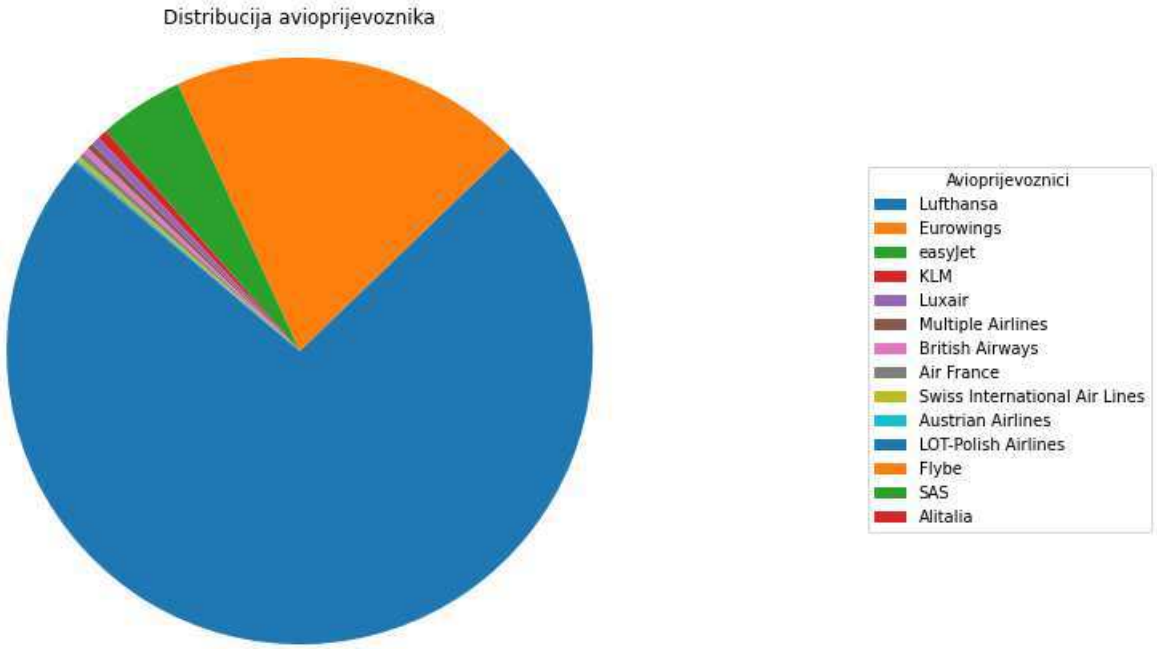


Slika 3.2. Histogram trajanja letova

Slika 3.2. prikazuje histogram iz kojeg je vidljivo da više od 30 000 letova što čini više od 50% cijelog skupa podataka traje manje od 5h dok dugih letova ima jako malo. Vidljivo je da postoji neravnoteža u podacima što može biti problem u predviđanju cijena. Stablo može postati pristrano prema predikcijama letova kraćeg trajanja jer su ti podaci dominantni. Drugi problem koji se može dogoditi je prekomjerno prilagođavanje (*engl. overfitting*) što može dovesti do složenijih stabala koja su teža za interpretaciju. Pretpostavka je da modeli neće biti dovoljno precizni u predviđanju cijena dugih letova zbog smanjenog broja podataka.

3.2. Ovisnost cijena letova o avioprijevoznicima

Poznata je činjenica da cijena avionske karte jako ovisi o avioprijevozniku. Tako postoje niskotarifne aviokompanije kao što su WizzAir i Ryanair i skuplje i luksuznije kao što je Turkish Airlines, Emirates i Qatar Airways. Zbog jake korelacije između aviokompanije i cijene bitno je vidjeti distribuciju podataka.

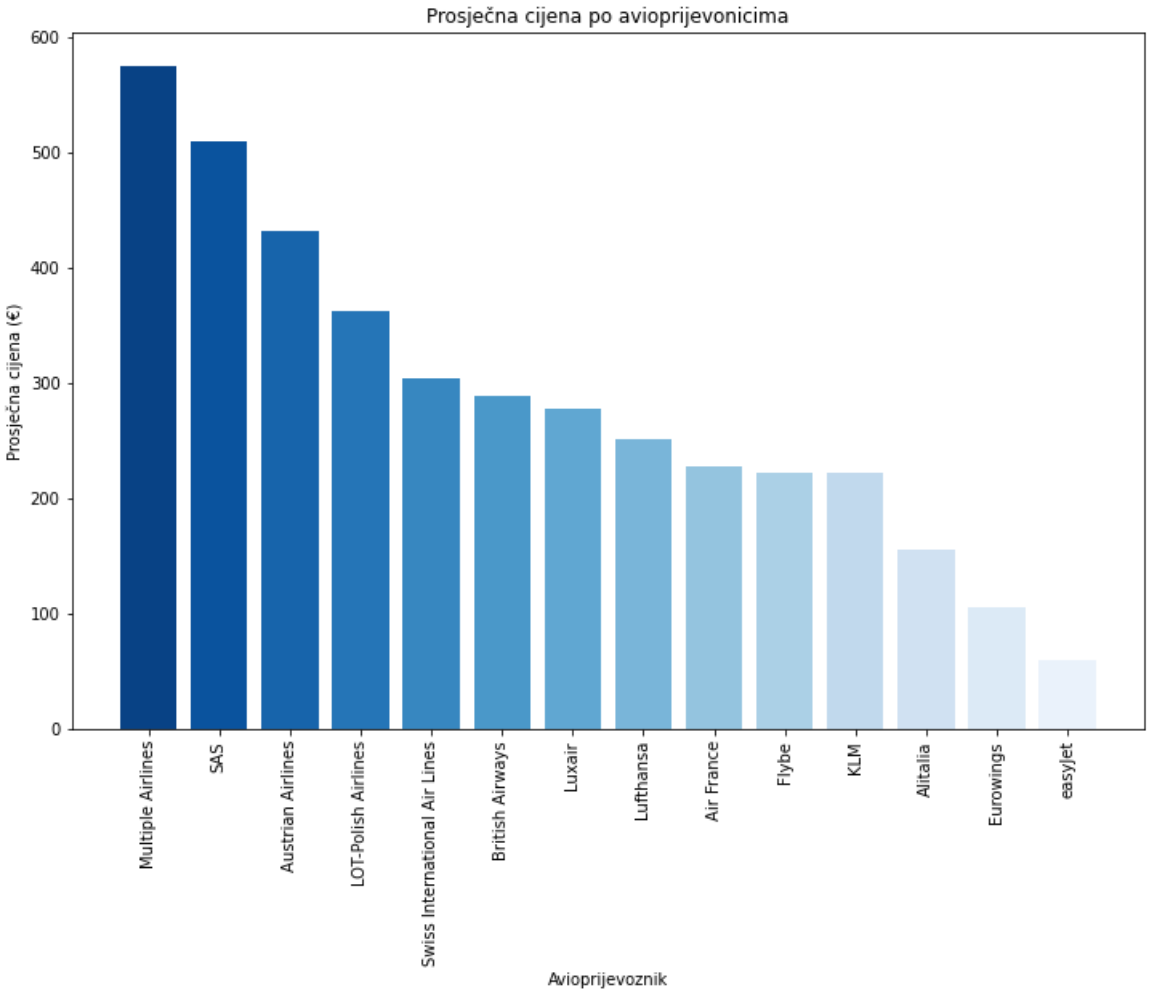


Slika 3.3. Raspodjela avioprijevoznika u skupu podataka

	Broj	**Postotak (%)**
Lufthansa	45912	73.311404
Eurowings	12289	19.622840
easyJet	2935	4.686552
KLM	341	0.544502
Luxair	290	0.463066
British Airways	197	0.314566
Air France	194	0.309775
Multiple Airlines	222	0.354485
Swiss International Air Lines	140	0.223549
Austrian Airlines	56	0.089420
LOT-Polish Airlines	44	0.070258
Flybe	3	0.004790
SAS	2	0.003194
Alitalia	1	0.001597

Tablica 3.1. Ukupan broj i postotak letova svakog avioprijevoznika

Slika 3.3 prikazuje tortni graf s raspodjelom avioprijevoznika u datasetu. Već je na prvi pogled vidljiva apsolutna dominacija Lufthanse i Eurowingsa te neravnomjernost podataka. Detaljniji prikaz s brojevima i postocima prikazan je u tablici 3.1. Te dvije aviokompanije čine više od 90% svih podataka u datasetu što će biti loše za generaliziranje i predviđanje cijena. Opet se događa isti slučaj kao s vremenima trajanja leta koja su isto bila neravnomjerno raspoređena i gdje je jedno vrijeme dominiralo.



Slika 3.4. Stupčasti graf prosječnih cijena letova za svaku aviokompaniju

Iz grafa sa slike 3.4 vidljivo je da je niskotarifna aviokompanija easyJet najjeftinija dok je najskuplje letjeti s različitim aviokompanijama. Detaljniji pregled podataka s brojčanim vrijednostima prikazan je u tablici 3.2.

Avioprijevoznik	Cijena (€)
Multiple Airlines	575.445946
SAS	510.000000
Austrian Airlines	431.232143
LOT-Polish Airlines	361.840909
Swiss International Air Lines	303.785714
British Airways	303.785714
Luxair	277.475862
Lufthansa	251.604069
Air France	227.587629
Flybe	222.000000
KLM	221.771261
Alitalia	156.000000
Eurowings	105.952966
easyJet	60.150596

Tablica 3.2. Ukupan broj i postotak letova svakog avioprijevoznika

Na temelju prikazanih podataka u tablici 3.2, koja sadrži prosječne cijene letova po avioprijevozniku, možemo donijeti nekoliko ključnih zaključaka koji se odnose na razlike u cijenama letova među različitim aviokompanijama. Ovi zaključci su dodatno podržani usporedbom s prethodnim grafovima koji prikazuju distribuciju trajanja letova i broj letova po aviokompaniji.

Kao što je navedeno na početku poglavlja 3.2. cijena leta ovisi o avioprijevozniku. Tako su letovi s različitim aviokompanijama puno skuplji od letova s istim aviokompanijama. Jedan od razloga za to je definitivno činjenica da su letovi s više presjedanja skuplji o čemu će biti riječ kasnije u sljedećim poglavljima.

Niskotarifni prijevoznici kao što su 'easyJet' i 'Eurowings' imaju niske prosječne cijene letova (60.15€ i 105.95€) što je u skladu s njihovim poslovnim modelom koji se temelji na pružanju osnovne usluge - leta i to po najnižim mogućim cijenama. Ti niskotarifni prijevoznici skupo naplaćuju bilo kakvu prtljagu i nažalost te cijene nisu uključene u

skup podataka pa zapravo nije moguće znati prave cijene za te niskotarifne letove. Najzastupljenija Lufthansa ima prosječnu cijenu letova 251.60€. S obzirom na to da većina podataka u tablici ima cijene oko 250€ pretpostavljamo da će modeli teže naučiti i ispravno predvidjeti letove s visokim cijenama.

3.3. Podaci o aerodromima

U današnje vrijeme skoro svaki veći grad ima aerodrom dok neki gradovi imaju više aerodroma. U tom slučaju su najčešće niskotarifni avioprijevoznici smješteni na jednom aerodromu dok su ostali smješteni na drugom aerodromu. Naš skup podataka sadrži informacije samo o njemačkim letovima i 19 njemačkih aerodroma.

Aerodrom	**Broj letova**	**Postotak (%)**
TXL Berlin-Tegel	13158	21.010443
DUS Düsseldorf	6506	10.388656
MUC München	6327	10.102833
HAM Hamburg	5978	9.545556
STR Stuttgart	4637	7.404273
CGN Köln/Bonn	4471	7.139027
NUE Nürnberg	3987	6.366365
FRA Frankfurt/Main	3855	6.155590
FDH Friedrichshafen	2651	4.233066
DRS Dresden	2617	4.178776
LEJ Leipzig/Halle	2275	3.632677
HAJ Hannover	1543	2.463333
BRE Bremen	1453	2.320123
FKB Karlsruhe/Baden-B	1284	2.050267
SCN Saarbrücken	793	1.266247
FMO Münster/Osnabrück	638	1.018874
DTM Dortmund	228	0.364066
PAD Paderborn/Lippsta	165	0.263469
RLG Rostock-Laage	60	0.095807

Tablica 3.3. Ukupan broj i postotak odlaznih letova za svaki aerodrom

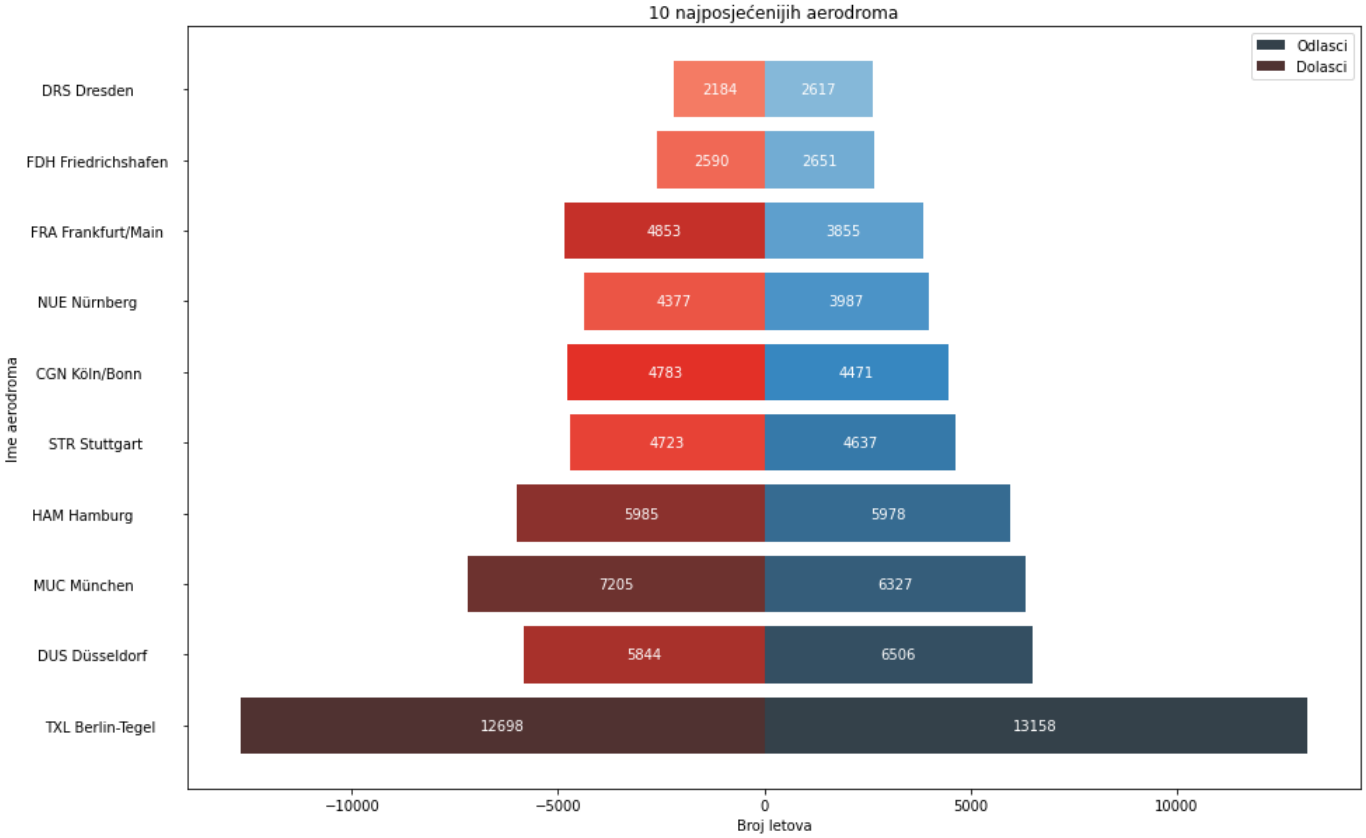
U tablici 3.3 prikazani su podaci i statistika odlaznih letova s pojedinih aeorodroma. Vidljivo je da su najčešći letovi iz Berlina, ali ima ih tek 21.01% . Slijede ga DUS Düsseldorf s 6506 letova (10.39%) i MUC München s 6327 letova (10.10%). Ostali aerodromi, kao što su HAM Hamburg, STR Stuttgart i CGN Köln/Bonn, imaju značajno manji udio letova, svaki s manje od 10% ukupnog broja letova. Manje prometni aerodromi, kao što su FMO Münster/Osnabrück, DTM Dortmund i RLG Rostock-Laage, imaju postotak letova manji od 1%.

Aerodrom	**Broj letova**	**Postotak (%)**
TXL Berlin-Tegel	12698	20.275924
MUC München	7205	11.504806
HAM Hamburg	5985	9.556734
DUS Düsseldorf	5844	9.331588
FRA Frankfurt/Main	4853	7.749178
CGN Köln/Bonn	4783	7.637403
STR Stuttgart	4723	7.541596
NUE Nürnberg	4377	6.989110
FDH Friedrichshafen	2590	4.135663
DRS Dresden	2184	3.487369
LEJ Leipzig/Halle	1935	3.089771
FKB Karlsruhe/Baden-B	1383	2.208348
HAJ Hannover	1280	2.043880
BRE Bremen	1108	1.769233
SCN Saarbrücken	963	1.537700
FMO Münster/Osnabrück	251	0.400792
DTM Dortmund	242	0.386421
PAD Paderborn/Lippsta	162	0.258679
RLG Rostock-Laage	60	0.095807

Tablica 3.4. Ukupan broj i postotak dolaznih letova za svaki aerodrom

Tablica 3.3 prikazuje ukupan broj i postotak dolaznih letova za svaki aerodrom. Podaci pokazuju da je aerodrom TXL Berlin-Tegel imao najviše letova (12698), što čini 20.28% ukupnog broja letova. Slijede MUC München s 7205 letova (11.50%), HAM Ham-

burg s 5985 letova (9.56%) i DUS Düsseldorf s 5844 letova (9.33%). S druge strane, aerodromi poput RLG Rostock-Laage i PAD Paderborn/Lippsta imali su znatno manje letova, s postotkom manjim od 0.1%.



Slika 3.5. 10 najposjećenijih aerodroma prema broju odlaznih i dolaznih letova

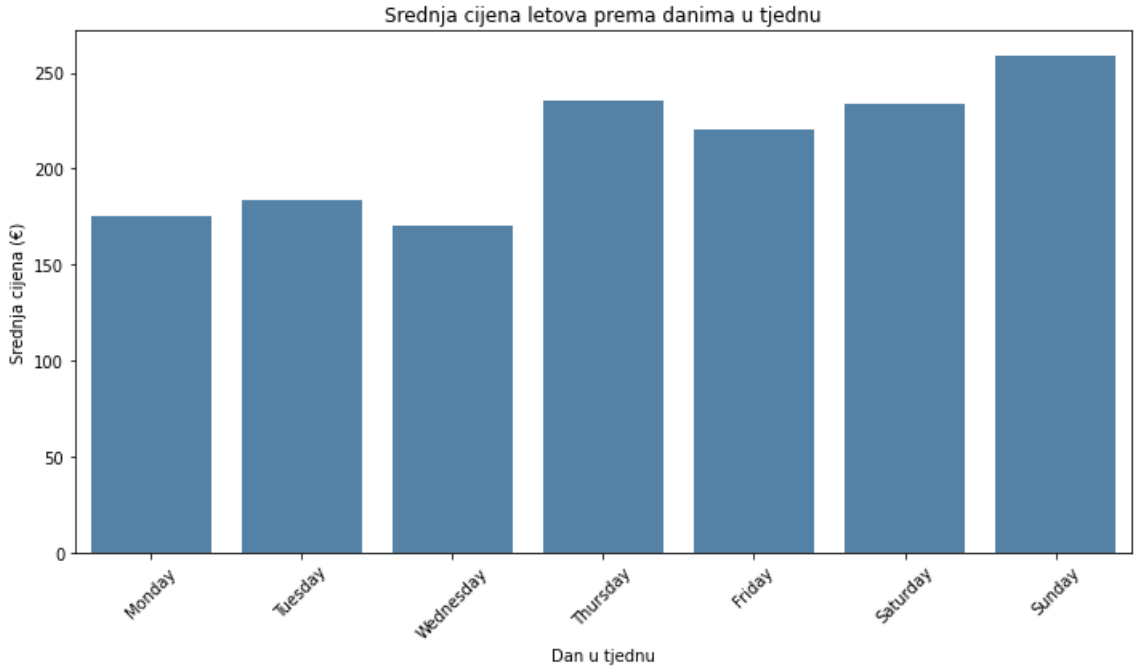
Slika 3.5 prikazuje top 10 aerodroma prema broju odlaznih i dolaznih letova. Na x - osi prikazan je broj letova s tim da pozitivne vrijednosti predstavljaju broj dolazaka, a negativne vrijednosti predstavljaju broj odlazaka s pojedinih aerodroma prikazanih na y - os. Vidljivo je da TXL Berlin-Tegel ima najveći broj letova s 12698 odlazaka i 13158 dolazaka kao što je već navedeno u tablicama 3.3. i 3.4. čime značajno dominira u prometu. Većina aerodroma ima sličan broj odlazaka i dolazaka, kao što je slučaj kod HAM Hamburg, STR Stuttgart, CGN Köln/Bonn i NUE Nürnberg.

Aerodromi s većim brojem letova, kao što su TXL Berlin-Tegel i MUC München mogu značajno utjecati na predikcije modela zbog svoje dominantne prisutnosti u podacima. S druge strane, aerodromi s manjim brojem letova mogu biti slabije zastupljeni i u modelu što može dovesti do pristranosti u predviđanju. Modeli stabala odlučivanja mogu imati

tendenciju favorizirati aerodrome s većim brojem podataka, što može smanjiti točnost predikcija za manje prometne aerodrome s manje dolaznih i dolaznih letova s dna tablica 3.3. i 3.4.

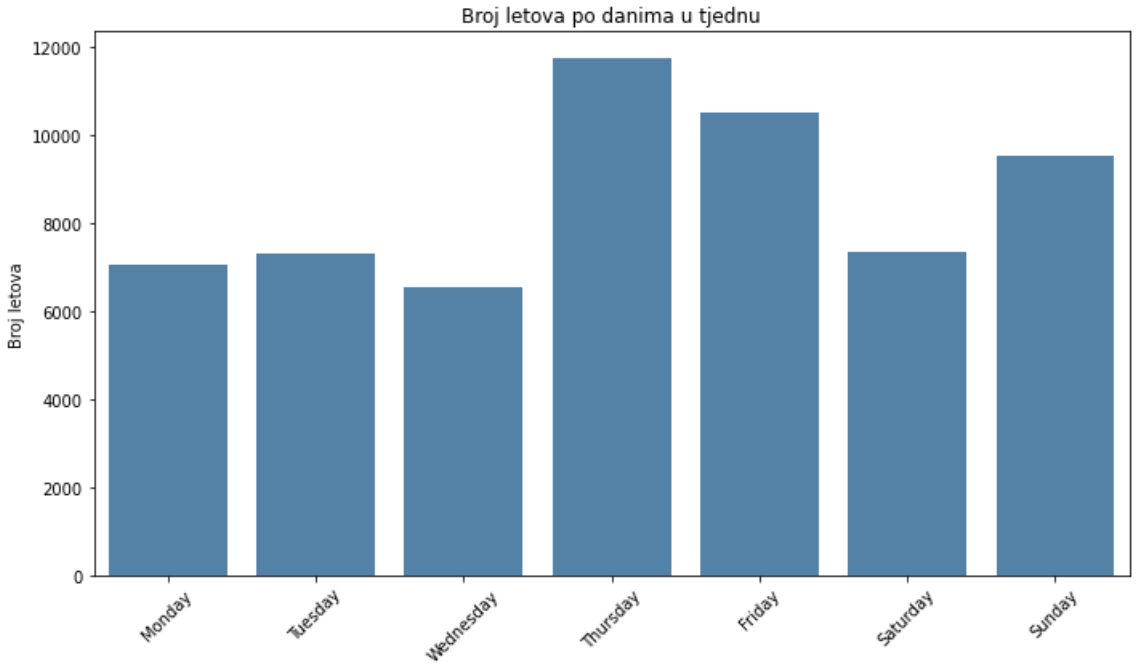
3.4. Ovisnost cijena letova o datumu i vremenu polaska

Vjerojatno prva stvar kod planiranja putovanja je odabir datuma putovanja (uz naravno destinaciju). Cijene na određene datume odnosno dane u tjednu mogu ljude "nagovoriti" da kupe avionsku kartu. Opće je poznato da su avionske karte jeftinije sredinom tjedna, dakle utorkom i srijedom dok su najskuplje avionske karte vikendom. U poglavlju će biti opisan odnos između cijena letova i dana u tjednu polaska aviona, broj letova po danima u tjednu te odnos cijena letova i satu polaska aviona.



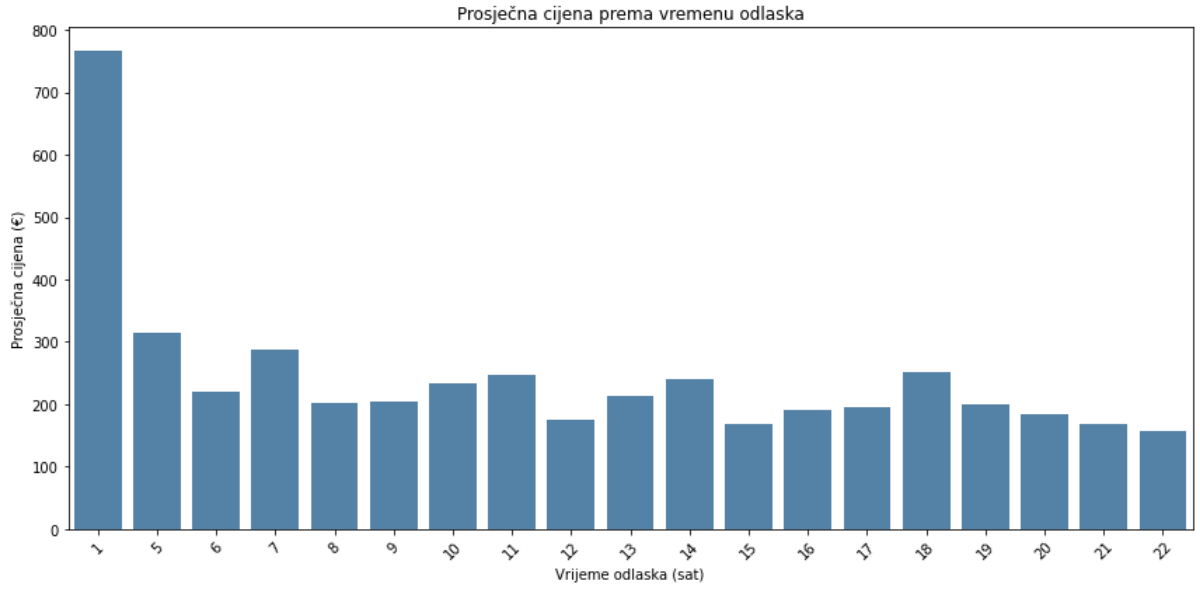
Slika 3.6. Stupčasti graf prosječnih cijena letova za svaki dan u tjednu

Slika 3.6 potvrđuje pretpostavku da su avionske karte najjeftinije početkom tjedna i to u srijedu dok su najskuplje u nedjelju. Međutim, srijedom ima najmanje letova u ponudi. Distribucija letova po danima u tjednu prikazana je ispod na Slici 3.6.



Slika 3.7. Stupčasti graf broja letova za svaki dan u tjednu

Pozitivna stvar u gore prikazanoj distribuciji je činjenica da nema jedne vrijednosti koja dominira nego svih podataka ima podjednako za razliku od prije navedenih distribucija. Time će pristranost stabla odlučivanja biti smanjena jer modeli imaju priliku učiti podjednako iz svih klasa što na kraju može povećati performanse modela i točnost predviđanja.

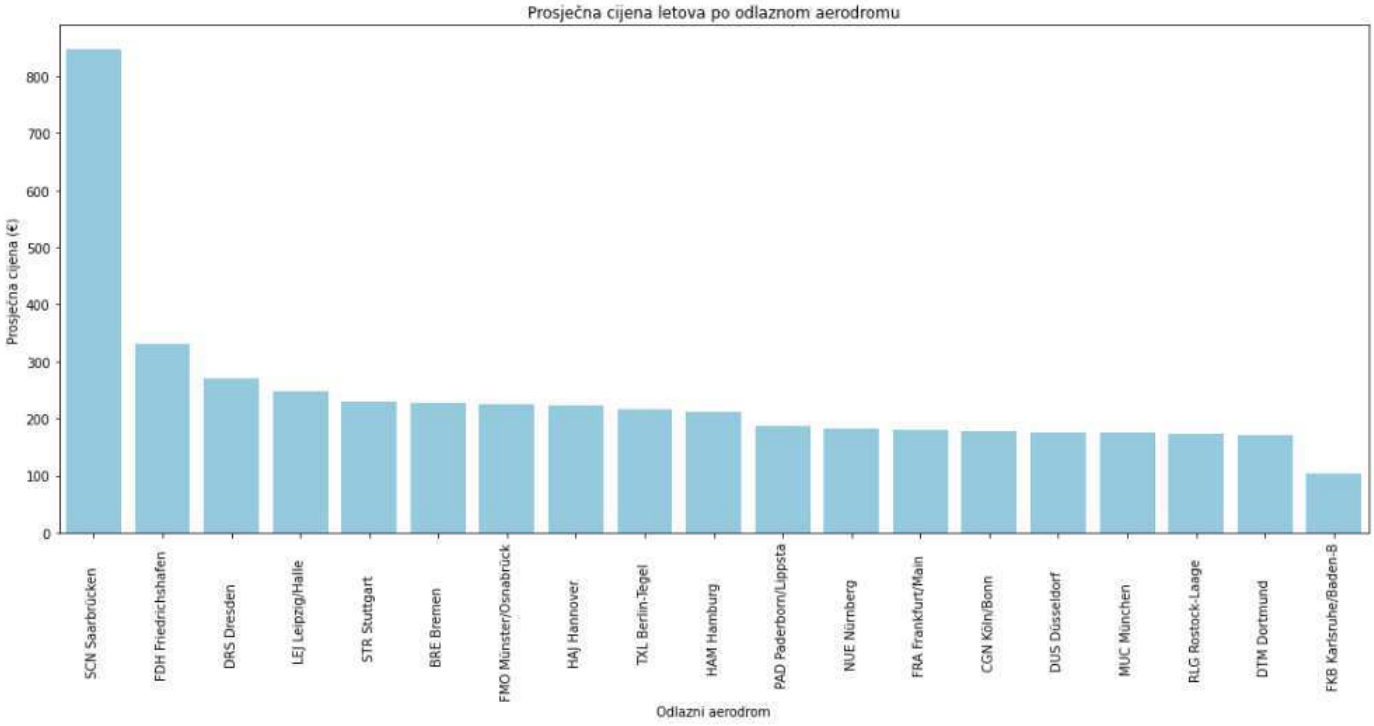


Slika 3.8. Stupčasti graf prosječne cijene i sata polaska

Iz slike 3.8 vidljivo je da postoji i ovisnost cijene o satu polaska i da su letovi u 1 ujutro najskuplji. Na x - osi prikazana su svi sati u kojima avion polazi dok y - os predstavlja prosječnu cijenu avionske karte. Detaljnijom analizom podataka došli smo do rezultata da u cijelom skupu podataka postoje samo 4 leta s vremenom polaska u 1 ujutro i da imaju 2 presjedanja zbog čega cijena i je toliko visoka. Ostale razlike u cijeni nisu pretjerano velike. Iz grafa je isto vidljivo da su cijene letova veće u jutarnjim satima nego u večernjim satima.

3.5. Ovisnost cijena letova o aerodromu

U poglavlju 3.3. opisani su odlazni i dolazni letovi njemačkih aerodroma. U ovom poglavlju bit će opisana ovisnost cijena letova s obzirom na odlazni i dolazni aerodrom. Ako grad ima više aerodroma letovi sa sporednog aerodroma na kojem posluju niskotarifne aviokompanije mogu biti daleko jeftiniji pa je bitno vidjeti koliko iznose cijene letove s pojedinih aerodroma.



Slika 3.9. Stupčasti graf prosječnih cijena odlaznih letova za svaki aerodrom

Slika 3.9. prikazuje prosječne cijene odlaznih letova za svaki aerodrom. Ovdje se ističe SCN Saarbrücken aerodrom s daleko najvišom cijenom i FKB Karlsruhe/Baden-B

aerodrom s najnižom cijenom. Ostali aerodromi imaju slične cijene.

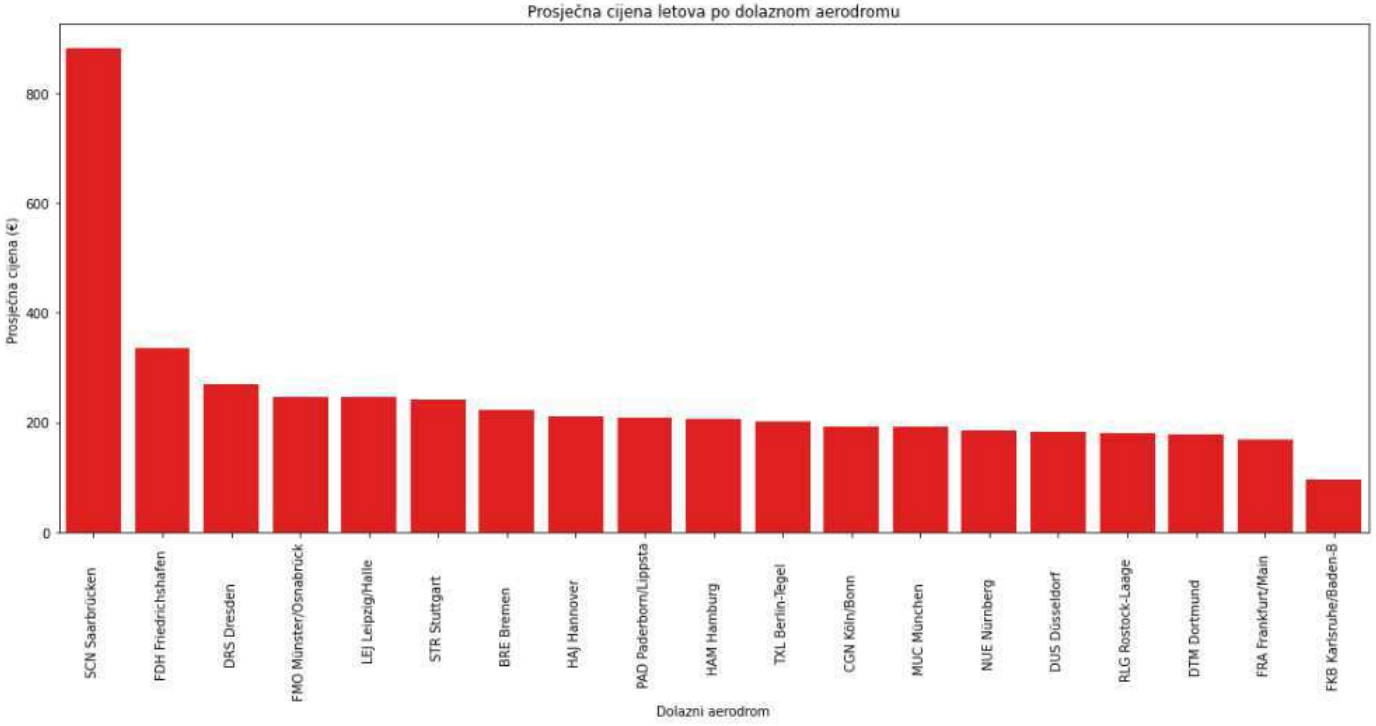
Sljedeća tablica prikazuje prosječne cijene letova za pri odlasku s pojedinih aerodroma.

Aerodrom	**Prosječna cijena (€)**
SCN Saarbrücken	847.0
FDH Friedrichshafen	331.0
DRS Dresden	271.0
LEJ Leipzig/Halle	248.0
STR Stuttgart	229.0
BRE Bremen	228.0
FMO Münster/Osnabrück	226.0
HAJ Hannover	224.0
TXL Berlin-Tegel	216.0
HAM Hamburg	212.0
PAD Paderborn/Lippsta	188.0
NUE Nürnberg	183.0
FRA Frankfurt/Main	181.0
CGN Köln/Bonn	178.0
DUS Düsseldorf	176.0
MUC München	175.0
RLG Rostock-Laage	174.0
DTM Dortmund	172.0
FKB Karlsruhe/Baden-B	103.0

Tablica 3.5. Prosječna cijena odlaznih letova za svaki aerodrom

Tablica prikazuje prosječne cijene odlaznih letova za različite aerodrome. Podaci pokazuju da SCN Saarbrücken ima najvišu prosječnu cijenu leta od 847 €, dok FKB Karlsruhe/Baden-B ima najnižu prosječnu cijenu leta od 103 €. Razlog najniže cijene na FKB Karlsruhe/Baden-B aerodromu je činjenica da tamo prometuju niskotarifne aviokompanije. Ove razlike u prosječnim cijenama mogu ukazivati na različite faktore kao što su udaljenost, broj dostupnih letova i konkurencija među avioprijevoznicima.

Dalje će biti opisane cijene dolaznih letova za različite aerodrome. Podaci prikazani na slici 3.10 su vrlo slični kao i za odlazne letove, SCN Saarbrücken se ističe s najvišim cijenama, a FKB Karlsruhe/Baden-B s najnižim. Iz prijašnjih tablica 3.3. i 3.4. vidljivo je da SCN Saarbrücken čini tek nešto više od 1% podataka tako da ove iznenadne visoke cijene ne bi trebale biti preveliki problem u budućnosti za predviđanje.



Slika 3.10. Stupčasti graf prosječnih cijena dolaznih letova za svaki aerodrom

Tablica 3.6 dosta je slična tablici 3.5., redoslijed se promijenio u par slučajeva, ali vidljivo je da su iznosi cijena dosta slični kao i kod odlaznih letova.

Smatram da odlazni i dolazni aerodromi nisu u direktnoj korelaciji s cijenom već da cijena više ovisi o dužini leta, broju presjedanja i aviokompaniji. Aerodromi su bitni zato što na njima prometuju visokotarifne ili niskotarifne aviokompanije i u ponudi imaju duže ili kraće letove o čemu cijena puno više ovisi.

Aerodrom	Prosječna cijena (€)
SCN Saarbrücken	881.0
FDH Friedrichshafen	335.0
DRS Dresden	269.0
FMO Münster/Osnabrück	247.0
LEJ Leipzig/Halle	245.0
STR Stuttgart	241.0
BRE Bremen	223.0
HAJ Hannover	211.0
PAD Paderborn/Lippsta	208.0
HAM Hamburg	207.0
TXL Berlin-Tegel	201.0
CGN Köln/Bonn	193.0
MUC München	191.0
NUE Nürnberg	185.0
DUS Düsseldorf	182.0
RLG Rostock-Laage	181.0
DTM Dortmund	177.0
FRA Frankfurt/Main	169.0
FKB Karlsruhe/Baden-B	96.0

Tablica 3.6. Prosječna cijena dolaznih letova za svaki aerodrom

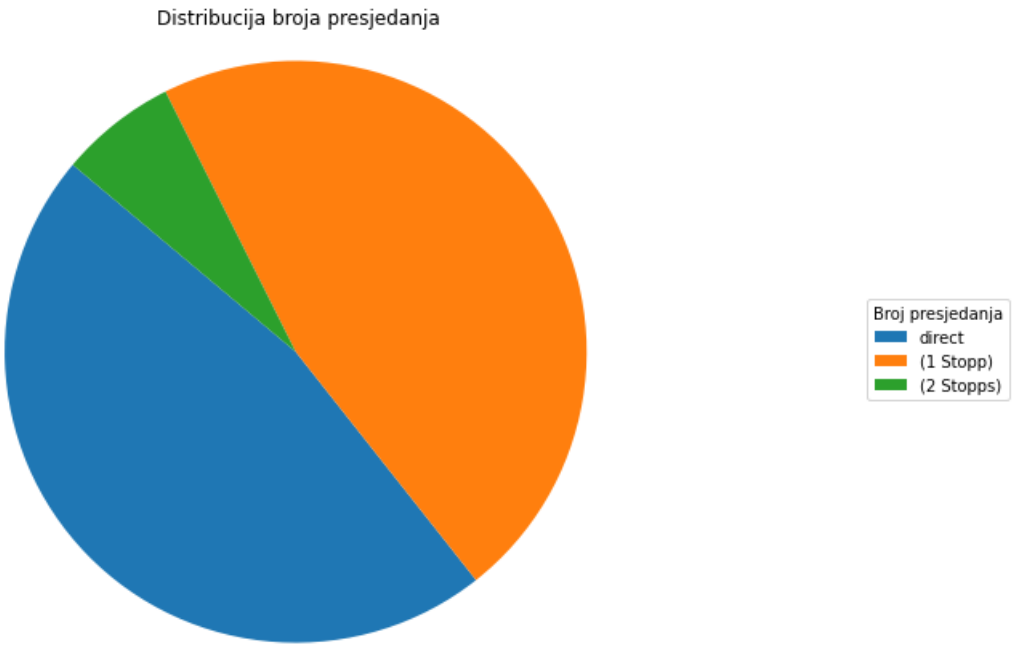
FKB Karlsruhe/Baden-B je opet najjeftiniji i jedini ima prosječnu cijenu manju od 100€. Iznenađujuće je da je FRA Frankfurt/Main drugi najjeftiniji aerodrom s obzirom na njegovu veličinu i činjenicu da je jedan od najprometnijih aerodroma na svijetu. Zato ne smije biti zaboravljeno da skup podataka sadrži informacije samo o letovima unutar Njemačke.

3.6. Ovisnost cijena letova o broju presjedanja

U analizi avionskih letova, važan faktor koji utječe na cijenu karte je broj presjedanja. Pretpostavka je da su direktni letovi jeftiniji nego letovi s presjedanjima, ali to naravno

ne mora biti pravilo. Direktni letovi mogu biti skuplji zbog udobnosti i uštede vremena putnicima. U ovom poglavlju bit će prikazani razni grafovi i ovisnosti cijena letova o broju presjedanja.

Bitno je vidjeti distribuciju podataka koja je prikazana na tortnom grafu na slici 3.11.

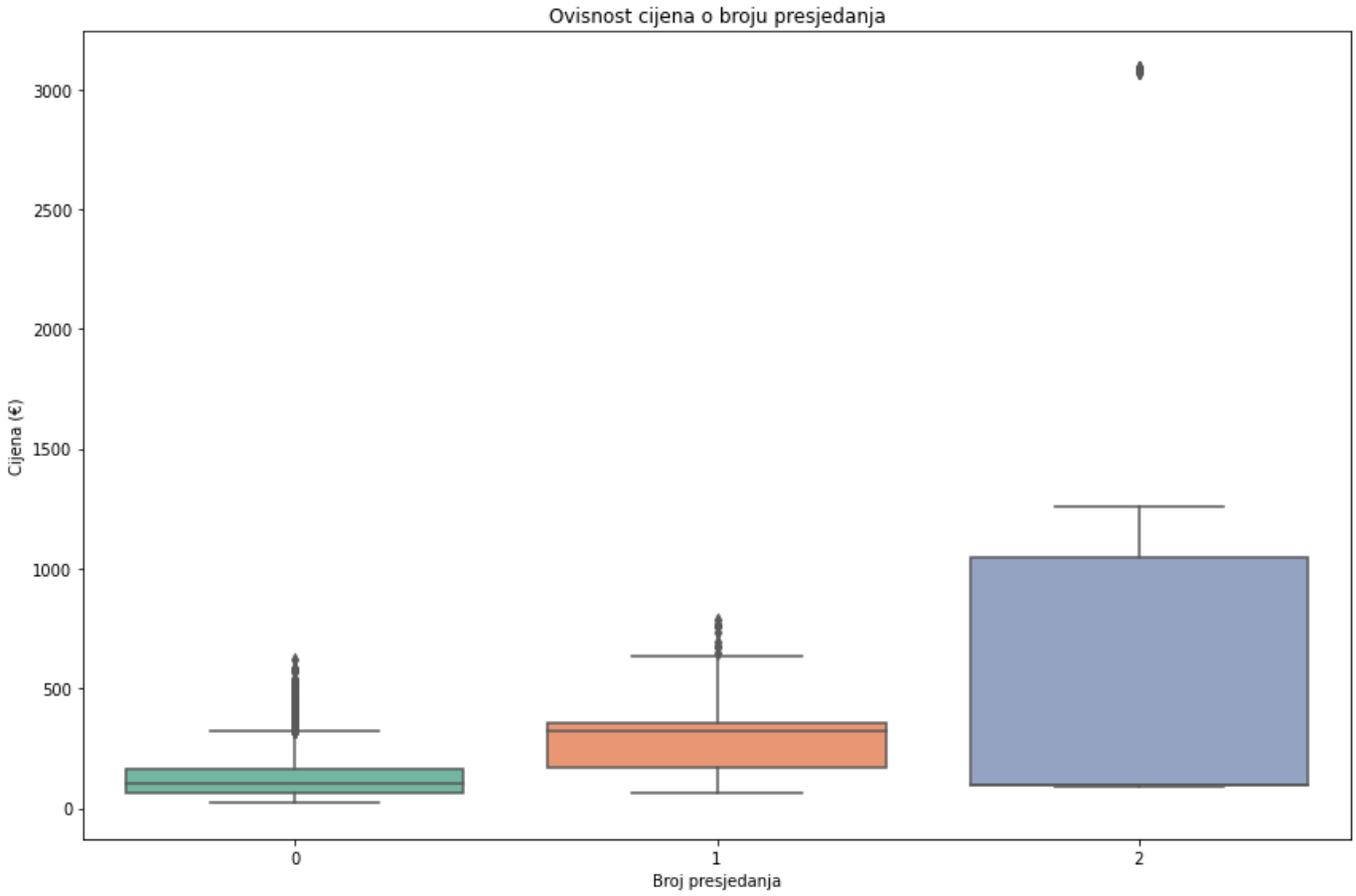


Slika 3.11. Distribucija broja presjedanja u skupu podataka

Tip leta	**Ukupno**
direkt	29278
1 presjedanje	29276
2 presjedanja	4072

Tablica 3.7. Ukupan broj letova s obzirom na broj presjedanja

Vidljivo je da ima podjednako direktnih letova i s jednim presjedanjem dok letova s dva presjedanja ima daleko najmanje. Razlog za to je to što se podaci odnose samo na njemačku i rijetko se događa da let ima više od jednog presjedanja.



Slika 3.12. Boxplot ovisnosti broja presjedanja o cijeni

Boxplot prikazuje raspodjelu cijena letova ovisno o broju presjedanja (0, 1 ili 2). Na x osi prikazan je broj presjedanja dok y os prikazuje cijenu letova. Stršeće vrijednosti prikazani su kao točke i predstavljaju značajno odstupanje od medijana. Iz boxplota je vidljivo da najniže cijene imaju direktni letovi dok i da su svi letovi ispod 1000€. Kod jednog presjedanja cijene su malo više i medijan im iznosi 320.0€, a cijene isto ne prelaze 1000€ dok u trećem slučaju kod letova s dva presjedanja postoje letovi za 3000€ koji su stršeće vrijednosti. Ti letovi imaju dosta veliki raspon cijena. Zanimljivo je da je medijan cijena s dva presjedanja 100.0€ dok je medijan cijena za direktne letove 107.0€. Zbog toliko velikih stršećih vrijednosti prosjek cijena nije dobra za analizu podataka pa su zato izračunati medijani koji su otporniji na stršeće vrijednosti. Medijan ili 50. percentil je srednja vrijednost po rednom broju u sortiranom nizu podataka pa tako vrijedi da 50% podataka ima veću vrijednost od medijana, a 50% podataka manju vrijednost od medijana.

3.7. Ovisnost cijena o vremenu kupnje avionske karte

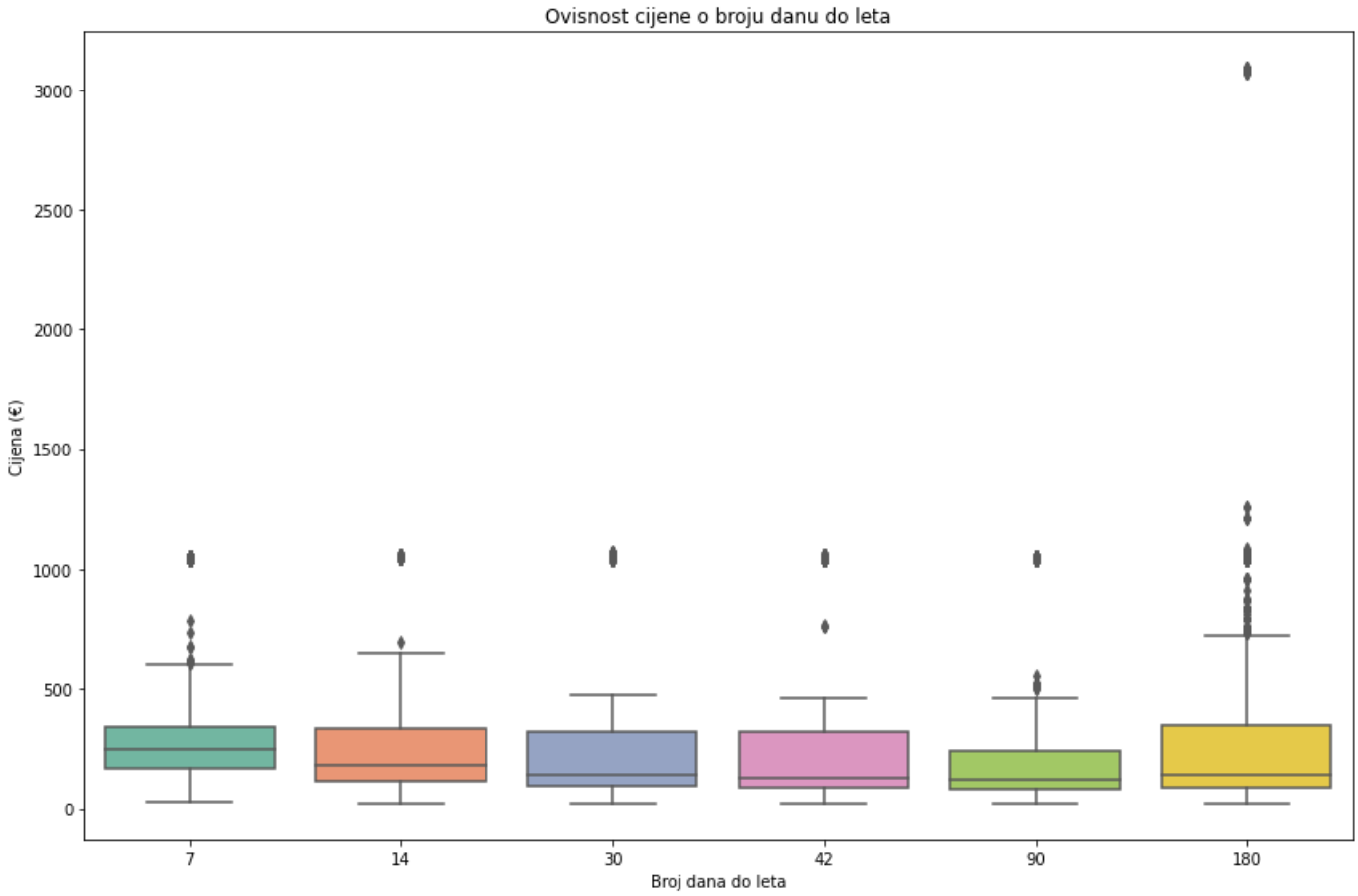
Jedna od bitnijih stvari kod kupovanja avionske karte je vrijeme kupnje. Kada je optimalno vrijeme kupnje? Treba li kupiti kartu tjedan dana prije leta ili pola godine prije leta? Ova pitanja često si postavlja svaki putnik i točan odgovor na ta pitanja zapravo ne postoji. Često se misli da je karta jeftinija ako se kupi puno ranije jer je tada potražnja manja, ali to nije pravilo. Kada se avionske karte kupuju par dana prije leta onda je potražnja velika i aviokompanije mogu povećati cijene.

U ovom poglavlju vidjet ćemo koje je najbolje vrijeme za kupnju avionske karte na temelju ponuđenih podataka za njemačke aviokompanije i letove. Skup podataka sadrži sljedeće informacije prikazane u tablici 3.8:

Broj dana prije leta	**Broj**	**Postotak (%)**
180	12672	20.234407
42	11222	17.919075
30	10092	16.114713
7	9949	15.886373
90	9748	15.565420
14	8943	14.280011

Tablica 3.8. Postoci i ukupni broj kupnji avionskih karata po danima prije leta

U tablici su svi podaci podjednako zastupljeni i ne postoji redak koji je skroz dominantan. Najviše podataka ima o cijenama avionskih karata 180 dana, tj. 6 mjeseci prije leta i oni čine 20.23% svih podataka dok je najmanje avionskih karata kupljeno 14 dana prije leta. Ti podaci čine 14.28% svih podataka.



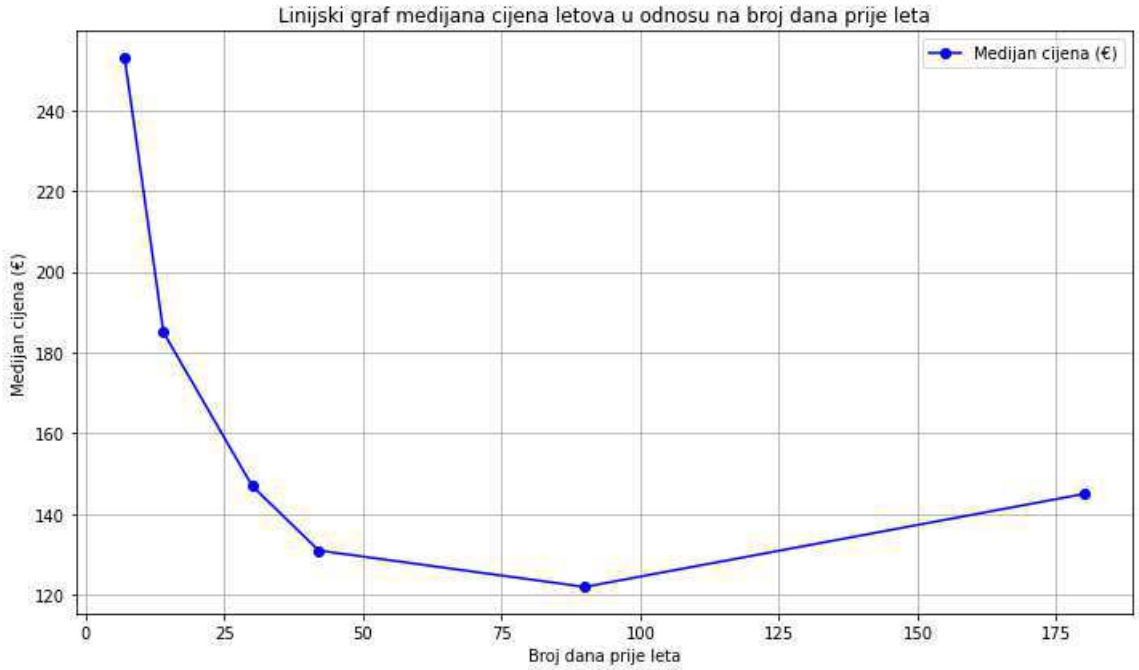
Slika 3.13. Boxplot cijena i broja dana do održavanja leta

U podacima opet ima dosta stršećih vrijednosti (*engl. outliers*) pa zbog toga nije optimalno računati prosječnu cijenu nego medijan kao u poglavlju 3.6. Boxplotovi su prikazani na slici 3.13 i na njima je vidljivo da su medijani dosta slični i da su rasponi cijena mali. Isto je vidljivo dosta stršećih vrijednosti u iznosu većem od 1000€.

Prave brojčane vrijednosti medijana su sljedeće:

Broj dana prije leta	**Cijena**
7	253.0
14	185.0
30	147.0
42	131.0
90	122.0
180	145.0

Tablica 3.9. Medijan cijena za broj dana do leta



Slika 3.14. Linijski graf ovisnost cijene avionske karte i broja dana prije leta

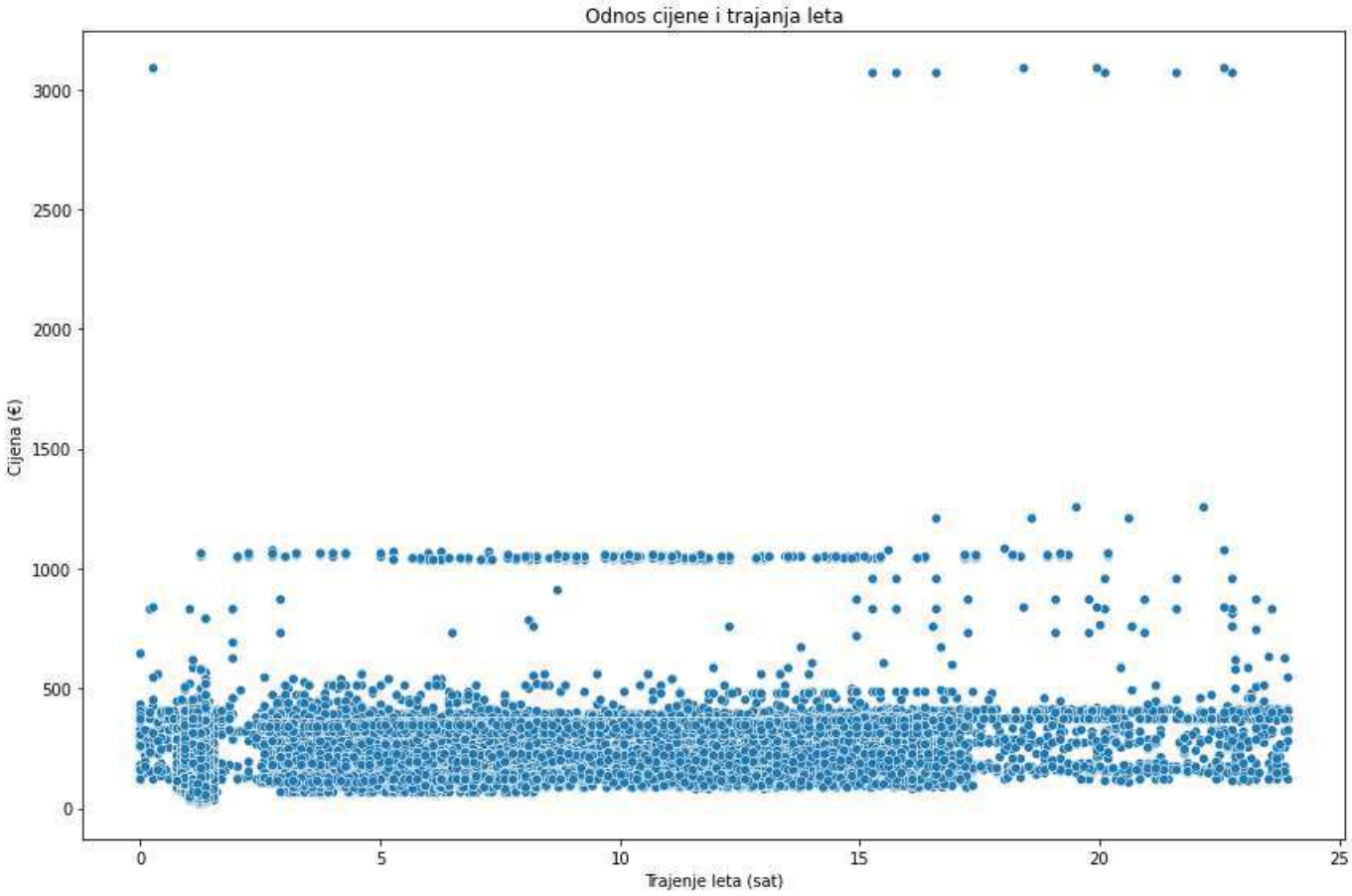
Linijski graf dao je puno informacija. Najoptimalnije vrijeme za kupiti avionsku kartu je 90 dana prije održavanja leta. Tada su cijene najniže. Najskuplje cijene avionskih karata su tjedan dana prije leta. Vidljivo je da je cijena niža što se karta ranije kupi i to "pravilo" vrijedi za prvih 90 dana. Pola godine, odnosno 180 dana prije leta cijena karata je veća i taj podatak je neočekivan. Trend pokazuje da će cijene karata od 7. do 90. dana prije leta padati dok će od 90. do 180. dana cijena rasti. Valja napomenuti da su linije u linijskom grafu zapravo samo aproksimacija jer podaci postoje samo za 7, 14, 30, 42, 90 i 180 dana prije polaska. S više podataka linijski graf bio bi precizniji i optimalno vrijeme kupnje bilo bi moguće preciznije odrediti.

Na temelju dobivenih podataka zaključeno je da je najbolje kupiti avionsku kartu 3 mjeseca prije polaska i da cijena uvelike ovisi o tome kada se avionska karta kupuje.

3.8. Ovisnost cijena letova o trajanju leta

Iako informacije o trajanju leta nisu bili zapisani u tablici podataka, trajanje leta vrlo je jednostavno za izračunati pomoću stupaca "departure_time" i "arrival_time" koji su prikazani na slici 3.1. Za pretpostaviti je da će duži letovi više koštati - bilo zbog veće udaljenosti pa aviokompanija mora trošiti više resursa ili zbog veće vjerojatnosti da pos-

toji potreba za presjedanjima za koje je rečeno da na cijenu ovise tako da je s brojem presjedanja cijena veća.



Slika 3.15. Točkasti graf ovisnosti dužine leta i cijene

Rezultati su prikazani pomoću točkastog grafa. Vidljivo je da su zapravo sve moguće varijante sadržane u skupu podataka i da se cijene većinom kreću do 1000€ uz stršeće vrijednosti iznosa 3000€. Na x osi prikazani su izračunata vremena trajanja leta dok y os sadrži cijene. Iz nacrtanog grafa nije moguće vidjeti ovisnost te dvije komponente. Cijene su u većini slučajeva manje od 500€ i za letove koji traju 1 sat i za letove koji traju više od 15 sati. Jedino što se može uočiti je veća koncentracija skupljih letova kada letovi traju duže.

Uočena je još jedna nepravilnost u podacima, a to je da postoje letovi kojima je trajanje 0 sati. To nije moguće tako da su ti podaci uklonjeni iz tablice. Isto tako uklonjene su stršeće vrijednosti s cijenama od 3000€ jer bi te vrijednosti loše utjecale na stabla odlučivanja.

4. Metrike i metode

Nakon detaljne analize svih podataka, brisanja stršećih vrijednosti i promatranja ovisnosti cijena o drugim parametrima vrijeme je za predikciju pomoću algoritama strojnog učenja koji se temelje na regresijskim stablima odlučivanja. U ovom poglavlju bit će opisane metrike koje su korištene, korišteni modeli i način skaliranja podataka.

4.1. Metrike procjene modela

Metrike za procjenu točnosti predviđanja koje su korištene su sljedeće:

- **Mean Absolute Error (MAE)**: MAE je prosjek apsolutnih razlika između predviđenih i stvarnih vrijednosti. Mjera je jednostavna za razumijevanje i interpretaciju jer predstavlja prosječnu pogrešku predikcija. Formula je sljedeća:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.1}$$

- **Mean Squared Error (MSE)**: MSE je prosjek kvadrata razlika između predviđenih i stvarnih vrijednosti. MSE kažnjava veće pogreške više nego MAE zbog kvadriranja razlika što ga čini korisnim za otkrivanje modela s velikim pogreškama. Kada model nema greške, MSE je jednak nuli. Kako se pogreška modela povećava, njegova vrijednost raste.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.2}$$

- **Root Mean Squared Error (RMSE)**: RMSE je kvadratni korijen MSE-a. Kao i MSE, RMSE kažnjava veće pogreške više. Mjeri prosječnu razliku između vrijed-

nosti predviđenih modelom i stvarnih vrijednosti. Pruža procjenu koliko dobro model može predvidjeti ciljanu vrijednost. Što je manja vrijednost srednje kvadratne pogreške, to je model bolji.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.3}$$

- **R^2 (koeficijent determinacije)**: R^2 mjera je koja prikazuje koliko dobro predviđene vrijednosti odgovaraju stvarnim vrijednostima. Vrijednost R^2 kreće se od 0 do 1, gdje veće vrijednosti označavaju bolje prilagodbu modela podacima.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.4}$$

- **Mean Absolute Percentage Error (MAPE)**: Srednja apsolutna postotna pogreška mjeri prosječnu veličinu pogreške koju proizvodi model ili koliko su daleko predviđanja u prosjeku.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{4.5}$$

4.2. Korišteni modeli

- **RandomForestRegressor**: *RandomForestRegressor* je ansambl metoda koja koristi više odluka stabala za regresiju. Svako stablo se trenira na različitim dijelovima podataka, a konačna predikcija se dobiva prosjekom predikcija svih stabala. Ovaj model smanjuje varijancu i sprječava pretreniranje, čime poboljšava točnost predikcija.

  Parametri koji se koriste kod RandomForestRegressor algoritma [9]:

  - `n_estimators`: broj stabala u šumi (default: 100).
  - `criterion`: funkcija za mjerenje kvalitete razdvajanja (default: `squared_error`).

- `max_depth`: maksimalna dubina stabla (default: None).
- `min_samples_split`: minimalni broj uzoraka potrebnih za podjelu unutarnjeg čvora (default: 2).
- `min_samples_leaf`: minimalni broj uzoraka potrebnih da bi list čvor bio valjan (default: 1).
- `min_weight_fraction_leaf`: minimalni udio težine uzoraka potreban da bi list čvor bio valjan (default: 0.0).
- `max_features`: broj značajki koje će se razmatrati pri traženju najboljeg razdvajanja (default: 1.0).
- `max_leaf_nodes`: maksimalni broj listova u stablu (default: None).
- `min_impurity_decrease`: minimalno smanjenje nečistoće potrebnog za podjelu čvora (default: 0.0).
- `bootstrap`: da li koristiti bootstrap uzorke pri izgradnji stabala (default: True).
- `oob_score`: da li koristiti out-of-bag uzorke za procjenu generalizacijske točnosti (default: False).
- `n_jobs`: broj radnih niti za paralelno računanje (default: None).
- `random_state`: kontrolira slučajnost stabala (default: None).
- `verbose`: kontrolira razinu ispisa (default: 0).
- `warm_start`: kada je postavljeno na True, ponovo koristi prethodno postavljena stabla i dodaje im nova (default: False).
- `ccp_alpha`: složenost parametra orezivanja minimalnog troška (default: 0.0).
- `max_samples`: ako je bootstrap=True, broj uzoraka za crtanje iz X za obuku svakog baznog procjenitelja (default: None).
- `monotonic_cst`: ograničenja monotonosti (default: None).

Sljedeći primjer pokazuje kako koristiti RandomForestRegressor u Pythonu:

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor

# Generiranje slucajnog skupa podataka
np.random.seed(42)
X = np.random.rand(100, 5)
y = np.random.rand(100)

# Podjela skupa podataka skup podataka za treniranje i testiranje
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Inicijalizacija modela RandomForestRegressor
model = RandomForestRegressor(random_state=42)

# Treniranje modela
model.fit(X_train, y_train)

# Predikcija na testnom skupu
y_pred = model.predict(X_test)

```

Listing 4..1: Primjer korištenja RandomForestRegressor algoritma

- **ExtraTreesRegressor**: *ExtraTreesRegressor* je također ansambl metoda koja koristi više odluka stabala, ali za razliku od *RandomForestRegressora*, koristi sve podatke za svaku podjelu i nasumično odabire točke podjele. Ovaj pristup smanjuje varijancu i dodatno smanjuje mogućnost pretreniranja. Jedina razlika u parametrima u odnosu na RandomForestRegressor je u parametru **verbose**. RandomForestRegressor koristi bootstrap uzorke pri izgradnji stabala (default: True) dok ExtraTreesRegressor ne koristi bootstrap uzorke (default: False). [10]

Jedina razlika u korištenju algoritma u kodu je u inicijalizaciji modela. Podjela podataka na skup za treniranje i testiranje je identičan. Model se inicijalizira na

sljedeći način:

```
from sklearn.ensemble import ExtraTreesRegressor
# Inicijalizacija modela ExtraTreesRegressor
model = RandomForestRegressor(random_state=42)

```

Listing 4..2: Primjer inicijalizacije ExtraTreesRegressor algoritma

- **DecisionTreeRegressor**: *DecisionTreeRegressor* koristi pojedinačno stablo odluke za regresiju. Model se gradi dijeljenjem podataka na sve manje skupove prema određenim pravilima dok se ne dođe do listova stabla. Iako je jednostavan za interpretaciju, ovaj model može patiti od pretreniranja, što smanjuje njegovu generalizaciju na neviđene podatke. Parametri ovog algoritma ne razlikuju se u odnosu na ExtraTreesRegressor i RandomForestRegressor osim što ih ima manje. [11] Parametri su sljedeći:
  - `criterion`: Funkcija za mjerenje kvalitete podjele. (default:`'squared_error'`).
  - `splitter`: Strategija za podjelu na svakom čvoru. Podržane opcije su `'best'` i `'random'` (default: `'best'`).
  - `max_depth`: Maksimalna dubina stabla (default: None).
  - `min_samples_split`: Minimalni broj uzoraka potrebnih za podjelu unutarnjeg čvora (default: 2).
  - `min_samples_leaf`: Minimalni broj uzoraka potrebnih da bi list čvor bio valjan (default: 1).
  - `min_weight_fraction_leaf`: Minimalni udio težine uzoraka potreban da bi list čvor bio valjan (default: 0.0).
  - `max_features`: Broj značajki koje će se razmatrati pri traženju najboljeg razdvajanja (default: None).
  - `random_state`: Kontrolira slučajnost (default: None).
  - `max_leaf_nodes`: Maksimalni broj listova u stablu (default: None).

- `min_impurity_decrease`: Minimalno smanjenje nečistoće potrebnog za podjelu čvora (default: 0.0).
- `ccp_alpha`: Složenost parametra koji se koristi za smanjenje minimalnog troška i složenosti (default: 0.0).
- `monotonic_cst`: Ograničenja monotonosti (default: None).

U odnosu na 4..1 jedina razlika u korištenju je također u inicijalizaciji modela i ona izgleda ovako:

```
from sklearn.ensemble import DecisionTreeRegressor
# Inicijalizacija modela DecisionTreeRegressor
model = DecisionTreeRegressor(random_state=42)

```

Listing 4..3: Primjer inicijalizacije DecisionTreeRegressor algoritma

4.3. Skaliranje podataka

U ovom poglavlju opisan je način skaliranje podataka pomoću *MinMaxScaler* transformacije podataka koja je korištena u procesu pripreme podataka za analizu i modeliranje. *MinMaxScaler* je metoda za skaliranje značajki tako da svaka značajka bude u zadanom rasponu, najčešće između vrijednosti 0 i 1.

MinMaxScaler iz *sklearn.preprocessing* modula transformira značajke tako da su skalirane pojedinačno kako bi bile unutar zadanog raspona na skupu podataka za treniranje. Transformacija se provodi prema sljedećim formulama:

$$X_{\text{std}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{4.6}$$

$$X_{\text{scaled}} = X_{\text{std}} \times (\text{max} - \text{min}) + \text{min} \tag{4.7}$$

gdje su min i max vrijednosti iz zadanog raspona *feature_range*.

Prednost *MinMaxScaler-a* su:

- **Jednostavnija interpretacija**: Skaliranje na raspon od 0 do 1 olakšava interpretaciju značajki jer su sve značajke u istom rasponu.
- **Veća učinkovitost za algoritme osjetljive na raspon podataka**: Neki algoritmi strojnog učenja mogu bolje funkcionirati kada su značajke unutar sličnog raspona (npr. linearna regresija).
- **Održavanje odnosa među podacima**: *MinMaxScaler* održava odnose među podacima budući da linearno skalira značajke.

U našem radu *MinMaxScaler* korišten je za skaliranje značajki kako bi bile unutar raspona od 0 do 1 prije nego što su podaci proslijeđeni modelima strojnog učenja. Time je osiguran jednak doprinos svake značajke u procesu treniranja modela i spriječeno je da značajke s većim numeričkim vrijednostima dominiraju nad ostalim značajkama. d Korištenje *MinMaxScaler* metode omogućilo je efikasnije treniranje modela strojnog učenja i poboljšalo točnost predviđanja u analizi kretanja cijena avionskih letova.

5. Rezultati

U ovom poglavlju bit će prikazani i rezultati evaluacije modela strojnog učenja korištenih za predikciju cijena avionskih letova. Korišteni modeli su RandomForestRegressor, ExtraTreesRegressor i DecisionTreeRegressor. Evaluacija modela provedena je korištenjem nekoliko metrike za procjenu točnosti: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 koeficijent determinacije i Mean Absolute Percentage Error (MAPE).

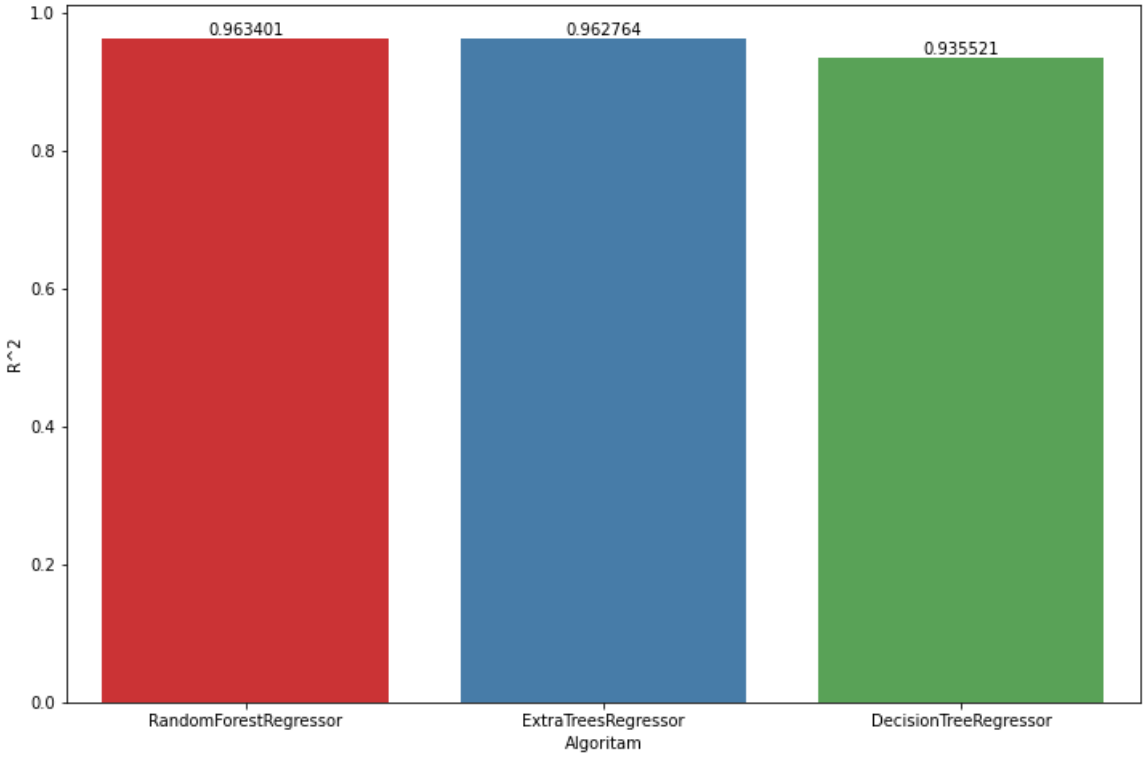
Model	**MAE**	**MSE**	**RMSE**	**R^2**	**MAPE**
RandomForestRegressor	17.15	1108.88	33.30	0.9634	11.70%
ExtraTreesRegressor	16.96	1128.17	33.59	0.9609	11.20%
DecisionTreeRegressor	17.78	1953.58	44.20	0.9355	11.79%

Tablica 5.1. Evaluacija modela strojnog učenja

Rezultati evaluacije prikazani u tablici 5.1. ukazuju na razlike u točnosti i performansama među korištenim modelima. Iz tablice je vidljiva velika sličnost u pogreškama kod RandomForestRegressor i ExtraTreesRegressor modela. RandomForestRegressor je pokazao najbolju točnost s najnižom vrijednošću RMSE (33.30) i najvišom vrijednošću R^2 (0.9634) što ukazuje na vrlo dobru prilagodbu modela podacima. Također je postigao nisku vrijednost MAE (17.15) i MAPE (11.70%) što znači da su predikcije modela u prosjeku bile vrlo blizu stvarnim vrijednostima. Vrlo slični rezultati ispali su i kod ExtraTreesRegressor algoritma. Vrijednosti RMSE (33.59) i R^2 (0.9609) su gotovo identične onima RandomForestRegressor-a, a vrijednosti MAE (16.96) i MAPE (11.20%), što ukazuje na vrlo dobru točnost predikcija. Vidljivo nižu točnost ima DecisionTreeRegressor u odnosu na prijašnja dva spomenuta algoritma. Vrijednost RMSE (44.20) je znatno viša, a R^2 (0.9355) je niži, što ukazuje na slabiju prilagodbu modela podacima. MAE (17.78) i

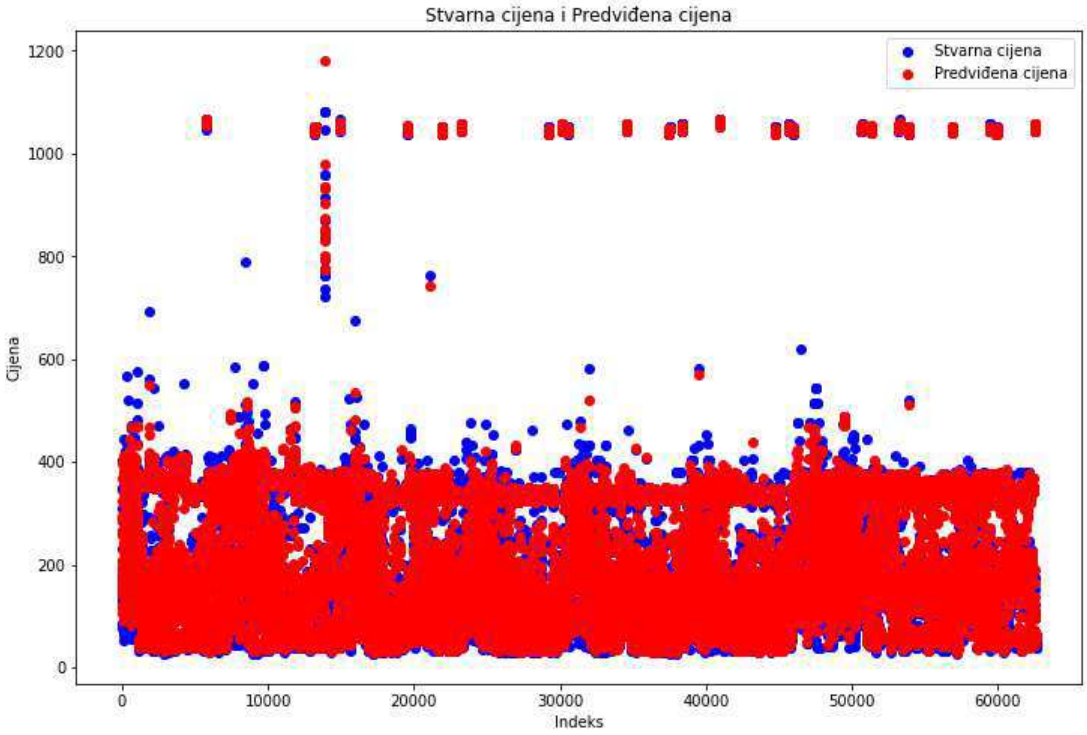
MAPE (11.79%) su također viši, što ukazuje na veće prosječne pogreške u predikcijama.

Na slici 5.1 prikazan je stupčasti graf koji vizualizira R^2 rezultate za sve tri modela. Vidljivo je da RandomForestRegressor i ExtraTreesRegressor postižu vrlo slične vrijednosti R^2, dok DecisionTreeRegressor zaostaje.



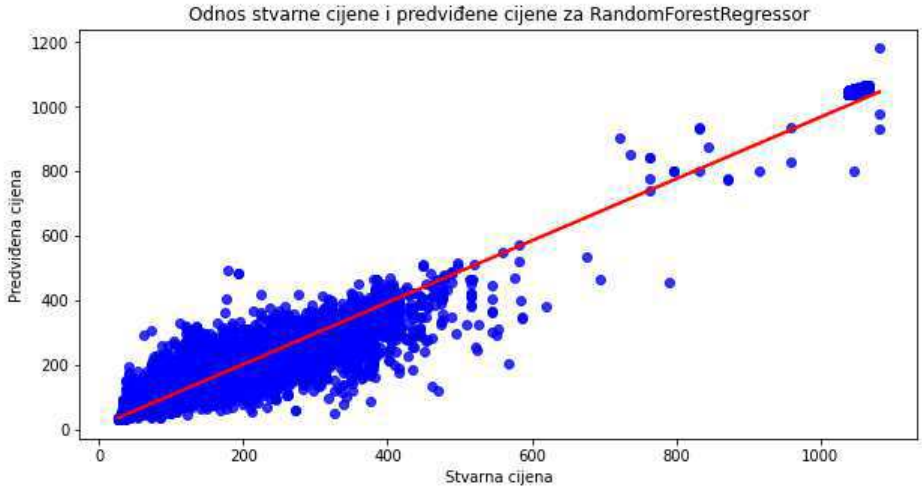
Slika 5.1. Stupčasti graf R^2 vrijednosti za sve algoritme

Ovi rezultati jasno pokazuju prednost korištenja ansambl metoda (RandomForestRegressor i ExtraTreesRegressor) nad pojedinačnim stablima odluke (DecisionTreeRegressor) za problem predikcije cijena avionskih letova.

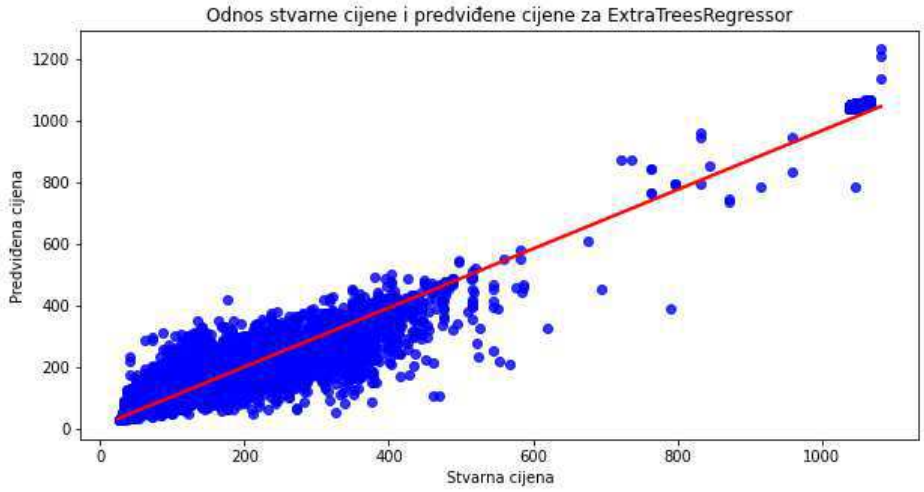


Slika 5.2. Odnos stvarne cijene i predviđene cijene za RandomForestRegressor u točkastom grafu

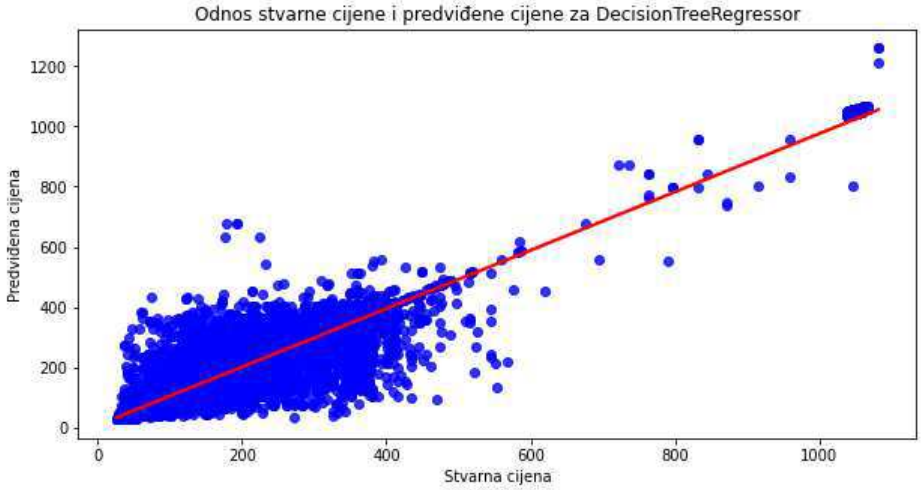
Slika 5.2. prikazuje predviđenu cijenu pomoću RandomForestRegressor-a i stvarnu cijenu za svaki let. Plave točke predstavljaju stvarnu cijenu, a crvene točke predstavljaju predviđenu cijenu. Na x-osi prikazan je redni broj leta za kojeg je cijena predviđena. Vidljivo je jako puno poklapanja i neka minijaturna odstupanja. Drugi i možda bolji prikaz rezultata i razlika između stvarne i predviđene cijene prikazan je na grafu ispod.



Slika 5.3. Odnos stvarne cijene i predviđene cijene za RandomForestRegressor



Slika 5.4. Odnos stvarne cijene i predviđene cijene za ExtraTreesRegressor



Slika 5.5. Odnos stvarne cijene i predviđene cijene za DecisionTreeRegressor

Grafovi prikazuje odnos između stvarnih i predviđenih cijena avionskih letova. Na horizontalnoj osi (x-osi) nalaze se stvarne cijene dok su na vertikalnoj osi (y-osi) prikazane predviđene cijene. Crveni pravac prikazuje najbolji slučaj kada je predviđena cijena jednaka stvarnoj cijeni.

Udaljenost točke od pravca je razlika između predviđene cijene i stvarne cijene. Dakle vrijedi pravilo da što je točka bliže pravcu to je predviđanje točnije. Veća raspršenost točaka oko pravca znači da je previđanje lošije i manje precizne. Iz grafa je vidljivo da su previđene cijene dosta blizu pravca, da ima dosta slučajeva gdje su predviđene cijene jednake stvarnim cijenama, ali i da ima nekih teških promašaja. Sve u svemu, na temelju srednjih pogrešaka i grafova može se reći da modeli dobro rade i da dosta dobro

predviđaju cijene.

Najbolji model prikazan je na slici 5.3 gdje su prikazani rezultati za algoritam RandomForestRegressor. Druga dva algoritma s malo slabijim rezultatima prikazani su na slikama 5.4 i 5.5. Iz grafova je vidljivo da su rezultati dosta slični kod ExtraTreesRegressora i RandomForestRegressora dok kod algoritma DecisionTreeRegressor vidljivo je da predviđene cijene puno više odstupaju od linijskog grafa što znači da je točnost modela puno manja. RandomForestRegressor je zato najbolji model za predviđanje cijena letova dok je DecisionTreeRegressor najgori model za predviđanje cijena letova

6. Zaključak

Cilj istraživanja bio je što točnije i preciznije predvidjeti cijene avionskih letova pomoću algoritama koji se temelje na regresijskim stablima odlučivanja, a rezultate za različite algoritme usporediti te grafički prikazati. U radu su korištena tri različita algoritma: RandomForestRegressor, ExtraTreesRegressor i DecisionTreeRegressor.

Prije samih predviđanja cijena bilo je potrebno urediti podatke što znači pretvoriti ih u dobar format (promjena stupaca vezanih za sate vremena polaska i dolaska), analizirati podatke, riješiti se stršećih vrijednosti koje mogu negativno utjecati na učenje modela i vidjeti o čemu sve cijena kao ciljna vrijednost ovisi. Podaci su u nekim stvarima bili prilično neravnomjerni. Tako na primjer su letovi aviokompanije Lufthansa činili čak 73.31% ukupnih podataka. Letovi s 2 presjedanja su bili jako malo zastupljeni u skupu podataka. To je naravno negativno utjecalo na učenje modela. Još jedna zanimljivost koja nije očekivana je ta da su cijene letova bile jeftinije ako se kupuju 3 mjeseca prije leta, a skuplje 6 mjeseci prije leta kao što prikazuje graf na slici 3.14. U podacima su bile neke stršeće vrijednosti s cijenama većim od 3000€. Isto tako postojali su letovi čije je trajanje bilo 0h. Nakon dobivenih rezultata koji nisu bili dobri odlučeno je da će se te stršeće vrijednosti i da će se reci s letovima koji traju 0 sati ukloniti iz skupa podataka. Nakon uklanjanja tih podataka rezultati su bili puno bolji.

Kod procjene točnosti modela korišteno je 5 različitih metrika: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) i R^2 (koeficijent determinacije).

Na temelju rezultata evaluacije može se zaključiti da je RandomForestRegressor najbolji model za predikciju cijena avionskih letova. Ovaj model pokazao je najbolju prilagodbu podacima s najvišim koeficijentom determinacije ($R^2 = 0.9634$) i najnižom sred-

njom kvadratnom pogreškom (RMSE = 33.30) i najnižom apsolutnom pogreškom (MSE = 1108.88). Iako je ExtraTreesRegressor postigao vrlo slične rezultate, RandomForestRegressor je malo bolji na kraju. Zato zaključujemo da je RandomForestRegressor najbolji model za ovaj problem.

S druge strane, DecisionTreeRegressor zaostaje za ansambl metodama u svim metrikama evaluacije što ukazuje na slabiju sposobnost generalizacije, veću sklonost pretreniranju i pretjeranoj sklonosti određenim podacima. Zbog toga se ne preporučuje kao prvi izbor za ovaj problem predviđanja cijena avionskih letova.

Iako su rezultati dobri i dosta precizni, uvijek postoji mjesta za napredak. Kako bi modeli bili bolji i precizniji potrebno je povećati broj podataka za treniranje i smanjiti dominantnost određenih podataka. Bitno je da ne postoji dominantnost podataka u stupcima kako modeli kako bi se modeli mogli učiti na svim tipovima podataka. Proširenje skupa podataka na letove iz cijelog svijeta umjesto korištenja samo njemačkih letova omogućili bi modelima da bolje uče i da na kraju bolje predviđaju cijene. Veći i raznovrsniji skup podataka može poboljšati generalizacijsku sposobnost modela.

Ovi modeli mogu biti korisni za putnike koji žele smanjiti troškove putovanja i koji nisu sigurni kada bi kupili avionsku kartu. Naravno, sva ova predviđanja ne treba uzimati kao 100% točne informacije jer zbog nekih vanjskih faktora cijene mogu naglo porasti i onda predviđanja cijena više ne vrijede. Međutim, to su najčešće neke izvanredne i nepredvidive situacije na koje ne možemo utjecati. Bez obzira na to konačni rezultati su dobri i modeli su se pokazali kao dobri za ovu vrstu problema.

Literatura

[1] K. Oglakcioglu, Y. S. Can, i F. Alagoz, "Prediction of optimal flight ticket purchase timing by using learning to rank methods", u *2023 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2023., str. 1–6. https://doi.org/10.1109/ASYU58738.2023.10296726

[2] K. Tziridis, T. Kalampokas, G. A. Papakostas, i K. I. Diamantaras, "Airfare prices prediction using machine learning techniques", u *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017., str. 1036–1039. https://doi.org/10.23919/EUSIPCO.2017.8081365

[3] L. Breiman, J. H. Friedman, R. A. Olshen, i C. J. Stone, *Classification and Regression Trees.* Wadsworth, 1984.

[4] "What is a Decision Tree? | IBM — ibm.com", https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes., [Accessed 16-06-2024].

[5] A. Meta, "A summary of classification and regression tree with application", 10 2016., str. 38–54. https://doi.org/10.33107/ubt-ic.2016.52

[6] L. Breiman, "Random forests", *Machine Learning*, sv. 45, br. 1, str. 5–32, 2001. https://doi.org/10.1023/A:1010933404324

[7] P. Geurts, D. Ernst, i L. Wehenkel, "Extremely randomized trees", *Machine Learning*, sv. 63, str. 3–42, 04 2006. https://doi.org/10.1007/s10994-006-6226-1

[8] F, Frederick, "German domestic air fares", 2021.

[9] “RandomForestRegressor — scikit-learn.org”, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html, [Accessed 27-06-2024].

[10] “ExtraTreesRegressor — scikit-learn.org”, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html, [Accessed 27-06-2024].

[11] “DecisionTreeRegressor — scikit-learn.org”, https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html, [Accessed 27-06-2024].

Sažetak

Primjena strojnog učenja u predviđanju cijena avionskih letova

Matej Lopotar

Cijene avionskih karata variraju prema mnogim faktorima kao što su dan leta, gradovi polaska i dolaska, trajanje leta, broj dana do leta, broj presjedanja i aviokompanija kojom se leti. Zrakoplovne kompanije određuju cijene na temelju algoritama koje su razvile. Testiranje i analiza provedeni su na skupu podataka koji opisuje njemačke letove. Cilj istraživanja je na temelju dostupnih podataka o letovima predvidjeti cijenu avionskih karata koristeći različite regresijske algoritme koji se temelje na stablima odlučivanja. Prije samog predviđanja podaci su detaljno analizirani i stršeće vrijednosti su uklonjene. Implementirana su tri regresijska algoritma: RandomForestRegressor, ExtraTreesRegressor i DecisionTreeRegressor. Rezultati su opisani pomoću sljedećih metrika: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 koeficijent determinacije i Mean Absolute Percentage Error (MAPE).

Ključne riječi: Stabla odlučivanja, Strojno učenje, Analiza podataka, Predviđanje cijena

Abstract

Application of Machine Learning in Prediction of Airfare Prices

Matej Lopotar

Airline ticket prices vary according to many factors such as flight day, departure and arrival cities, flight duration, number of days until the flight, number of stopovers, and the airline company. Airline companies determine prices based on algorithms they have developed. Testing and analysis were conducted on a dataset describing German flights. The aim of the research is to predict airline ticket prices based on available flight data using various regression algorithms based on decision trees. Before the prediction itself, the data was analyzed and outliers were removed. Three regression algorithms were implemented: RandomForestRegressor, ExtraTreesRegressor, and DecisionTreeRegressor. The results are described using the following metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 coefficient of determination, and Mean Absolute Percentage Error (MAPE).

Keywords: Decision Trees, Machine Learning, Data Analysis, Price prediction