

UNIVERZITET U BEOGRADU
МАТЕМАТИЧКИ ФАКУЛТЕТ



Andelka Đ. Milovanović

**PRIMENA METODA MAŠINSKOG
УЧЕЊА ЗА ПРЕДVIĐАЊЕ ПОТРАŽНJE
AUTOMOBILSKIH REZERVNIХ ДЕЛОВА**

master rad

Beograd, 2021.

Mentor:

dr Aleksandar KARTELJ, docent
Univerzitet u Beogradu, Matematički fakultet

Članovi komisije:

dr Vladimir FILIPOVIĆ, redovan profesor
Univerzitet u Beogradu, Matematički fakultet

dr Milan BANKOVIĆ, docent
Univerzitet u Beogradu, Matematički fakultet

Datum odrbrane: _____

Hvala ljudima sa severa (M&S) za svaki vid podrške.

*Hvala profesoru Aleksandru Kartelju.
Hvala Slobodanu, Isidori i Dari.*

Naslov master rada: Primena metoda mašinskog učenja za predviđanje potražnje automobilskih rezervnih delova

Rezime: Predviđanje potražnje (*eng. demand forecasting*) predstavlja izazovan i veoma zastupljen problem u svakodnevnom industrijskom i biznis svetu. Glavni ishod dobrog predviđanja potražnje je optimalna ušteda resursa, a samim tim i povećanje doprinosu na razne načine. Cilj ovog rada je predviđanje potražnje automobilskih delova nad realnim podacima prikupljenim iz jedne kompanije u Švedskoj. Podaci predstavljaju ograničene informacije vezane za rezervacije popravki automobila raznih marki, na različitim lokacijama, dobijene isključivo putem zakazivanja preko internet portala. Rad analizira prikupljene podatke na nekoliko načina. Prva analiza posmatra podatke predstavljene u vidu vremenske serije, istraživajući dnevni i nedeljni slučaj kroz tri metode: Arima, Prophet i XGBoost. Druga analiza se bavi istraživanjem svih prikupljenih podataka, odvajanjem bitnih atributa, njihovim encodiranjem i regresionim problemom predviđanja broja rezervacija nad nekoliko nivoa granularnosti podataka.

Ključne reči: vremenske serije, mašinsko učenje, predviđanje potražnje, automobilска industrija, arima, prophet, xgboost

Sadržaj

| | |
|--|-----------|
| 1 Uvod | 1 |
| 1.1 Motivacija | 1 |
| 1.2 Opis problema | 2 |
| 1.3 Pregled dosadašnjih istraživanja | 3 |
| 2 Opis korišćenih metoda i metrika | 5 |
| 2.1 ARIMA | 5 |
| 2.2 Prophet | 6 |
| 2.3 XGBoost | 7 |
| 2.4 Metrike | 8 |
| 2.5 Impact Encoding | 9 |
| 3 Prikaz rada metoda i rezultati | 11 |
| 3.1 Podaci | 11 |
| 3.2 Dnevni nivo - na nivou države | 13 |
| 3.3 Nedeljni nivo - na nivou države | 23 |
| 3.4 XGBoost veća granularnost - na nivou automehaničarskih radnji i marki automobila | 26 |
| 3.5 Diskusija rezultata | 29 |
| 4 Zaključak i pravci daljeg rada | 31 |
| Bibliografija | 33 |

Glava 1

Uvod

1.1 Motivacija

Problem predviđanja potražnje (*eng. demand forecasting*) predstavlja dosta bitan i zastupljen problem u svim industrijama. On predstavlja najbolju procenu kakva će potražnja za nekim proizvodom/proizvodima biti u budućnosti, pod nekim datim ulaznim parametrima odnosno pretpostavkama [8]. Ishod optimalnog procesa predviđanja zavisi od prirode biznisa. U knjizi [8], autor objašnjava kako procesi predviđanja zavise od kompanije do kompanije i nešto najbolje za jednu, ne mora biti (a uglavnom i nije) najbolje za drugu kompaniju. Razlike se ogledaju u dostupnim podacima, proizvodima, ciljevima, mušterijama i slično. U radu [16] je predstavljen kratak pregled istraživanja na temu predviđanja u raznim sferama: potražnje putovanja avionom, potražnje automobila, potražnje električne energije u Maleziji, odakle se može zaključiti široka zastupljenost ovog problema.

Predviđanje potražnje je jedna od bitnijih tema kada se radi o optimizaciji. Optimizaciji vezanoj za uštedu energije, uštedu materijala, smanjenje troškova, povećanje efikasnosti, donošenje odluka na nivou kompanija, gradova, država. Indirektno je vezano za uštedu resursa i održiv razvoj društva i sredine [14]. Takođe, sve navedeno čini glavne pokretače tranzicije ka društvu sa manjim nivoom ugnjenika (*eng. low carbon society*) [3] i zastupljeno je u industriji popravke automobila, koja je pre svega važna zbog teme ovog rada.

1.2 Opis problema

Kao što je pomenuto, tema ovog rada je vezana za automobilsku industriju. Konkretnije, za industriju koja se bavi nabavkom i distribuiranjem delova, kao i raznim popravkama automobila. Problem je kroz rad posmatran na dva načina:

1. kao vremenska serija
2. kao metod zasnovan na atributima (*eng. feature-based*)

Kao što je pomenuto u [2], predviđanje potražnje se može podeliti na kvantitativno i kvalitativno. Unutar kvantitativnih metoda nalaze se: vremenske serije, metode zasnovane na atributima i hibridne metode. Vremenske serije koriste samo istorijske podatke da predvide budućnost. Metode zasnovane na atributima koriste promenljive koje opisuju problem i procenjuju potražnju kao linearnu/nelinearnu ili nekakvu sličnu funkciju promenljivih. Hibridne metode mešaju jednu ili više tipova tehnika i njima se u ovom radu nećemo baviti.

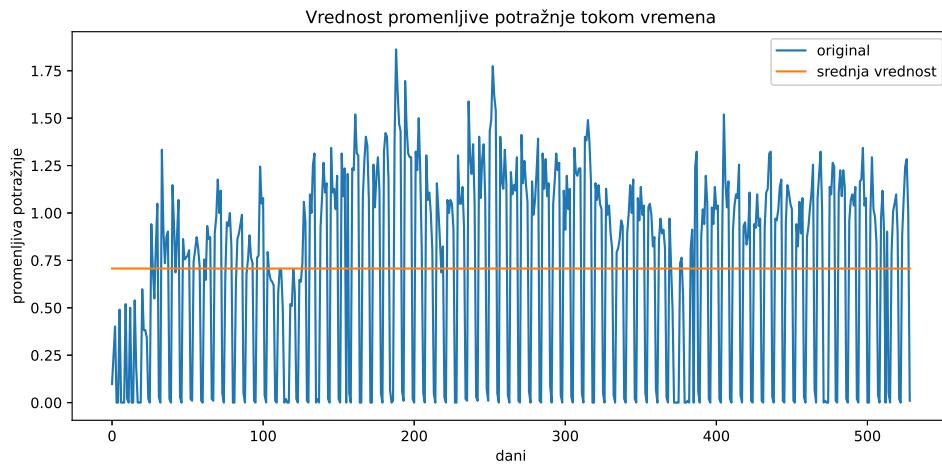
Problem kao vremenska serija

Vremenska serija je niz vrednosti $x_1, x_2, x_3, \dots, x_t$ koji je uzorkovan sekvencijalno u zavisnosti od promenljive koja je vreme. Vremenska serija se može agregirati, u zavisnosti od koraka uzorkovanja [5]. U ovom radu, originalni istorijski podaci promenljive koja predstavlja potražnju su dobijeni na dnevnom nivou, a agregacija je rađena za nedeljni nivo sa ciljem ispitivanja. Cilj predviđanja vremenske serije je izračunavanje vrednosti u nekom trenutku budućnosti x_{t+h} , ako su poznate vrednosti x_1, x_2, \dots, x_t . U zavisnosti od vrednosti h , postoji predvidanje jedan korak u napred i više koraka u napred. U ovom radu ispitivana vremenska serija je predstavljena kao serija jedne promenljive (*eng. univariate forecasts*). Na grafiku 1.1 se može videti primer jedne vremenske serije, a opis metode rešavanja ovakvog problema biće predstavljen u sledećoj sekciji.

Problem posmatran kroz attribute

Kako razumevanje problema u vidu vremenske serije nije dalo značajne rezultate na većim granularnostima podataka, problem je posmatran kao regresioni problem koji zavisi od nekoliko atributa koji ga opisuju i rešavan je ansambl metodom ekstremnog gradijentnog pojačavanja. Cilj je bio pronaći vezu između dostup-

GLAVA 1. UVOD



Slika 1.1: Primer vremenske serije iz rada

nih atributa i ciljne promenljive koja predstavlja vrednost potražnje. Primer skupa atributa i ciljne promenljive prikazan je na grafiku 1.2, a opis metode rešavanja ovakvog problema biće predstavljen u sledećoj sekciji.

| | vehicle_make | garage | reachable_population | year | week_of_year | number_of_competitors | x_unit_cost | x_units | demand_value | is_training_data |
|-------|--------------|--------|----------------------|------|--------------|-----------------------|-------------|----------|--------------|------------------|
| 0 | AE | CU | 0.557521 | 2020 | 48 | 0.218750 | 0.741525 | 2.100000 | 0.666667 | True |
| 1 | AE | CX | 0.701217 | 2020 | 39 | 0.203125 | 0.595870 | 1.966667 | 0.666667 | True |
| 2 | AE | DC | 1.835690 | 2020 | 38 | 1.578125 | 1.324153 | 0.366667 | 0.666667 | True |
| 3 | AE | DD | 1.089601 | 2020 | 45 | 1.375000 | 1.101695 | 2.133333 | 1.333333 | True |
| 4 | AE | DT | 1.821892 | 2019 | 52 | 1.609375 | 1.191737 | 3.033333 | 0.666667 | True |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 25799 | BI | CS | 0.701217 | 2021 | 8 | 0.203125 | 0.825212 | 6.600000 | 3.333333 | True |
| 25800 | BI | CS | 0.701217 | 2021 | 9 | 0.203125 | 1.006356 | 2.333333 | 2.000000 | True |
| 25801 | BI | CS | 0.701217 | 2021 | 10 | 0.203125 | 1.006356 | 1.666667 | 1.333333 | True |
| 25802 | BI | CS | 0.701217 | 2021 | 11 | 0.203125 | 1.006356 | 1.666667 | 1.333333 | True |
| 25803 | BI | CS | 0.701217 | 2021 | 12 | 0.203125 | 1.006356 | 1.000000 | 0.666667 | False |

25804 rows × 10 columns

Slika 1.2: Primer skupa atributa i ciljne promenljive iz rada

1.3 Pregled dosadašnjih istraživanja

Podaci korišćeni u radu dolaze iz novog izvora (pokrenut krajem 2019. godine), koji je dostupan korisnicima za rezervisanje popravki automobila, tako da su istraživanja nad ovim podacima prilično mlada i ne javno dostupna. Što se tiče same

GLAVA 1. UVOD

automobilske industrije i predviđanja potražnje vezane za popravke automobila i rezervne delove, takva istraživanja postoje. U radu [4] na stranicama 18-22, dat je veoma detaljan pregled različitih metoda koje se koriste za predviđanja kao što su: Single Exponential Smoothing, Croston's Method, Moving Average, ARIMA, Neural Network... kao i detaljan tabelarni pregled autora koji su koristili neke od metoda u svojim istraživanjima. U master radu [5], prikazan je opširan pristup predviđanju potražnje automobilskih rezervnih delova dubokim učenjem.

Rad [17] se bavi predviđanjem prodaje proizvoda u maloprodajnoj industriji nad realnim podacima, korišćenjem Prophet alata za predviđanje kod vremenskih serija. U radu [11] je predstavljeno predviđanje prodaje rezervnih delova ARIMA metodom. Navedeni pregled istraživanja potvrđuje koliko je problem predviđanja potražnje zastupljen u svakodnevnom svetu i pokazuje širinu isprobanih metoda koje mogu biti iskorišćene za rešavanje problema. Neke od navedenih metoda će biti isprobane u ovom radu.

Glava 2

Opis korišćenih metoda i metrika

U ovom poglavlju biće dat kratak pregled korišćenih metoda za rešavanje definisanog problema (videti poglavlje 1.2) predviđanja potražnje u industriji popravke automobila. Te metode su:

1. ARIMA
2. Prophet
3. XGBoost

Pored korišćenih metoda, ukratko će biti opisane i metrike za evaluaciju modela i nekoliko pojmovna značajnih za rad.

2.1 ARIMA

ARIMA (*eng.* Autoregressive Integrated Moving Average) predstavlja algoritam za predviđanje ciljne promenljive kod vremenskih serija. Ona je dizajnirana da identificuje obrasce u potražnji koja se desila i obrasce koji se ponavljaju kroz vreme [8]. Takođe, ARIMA je model univarijantne vremenske serije, gde je predviđanje zasnovano isključivo na prethodnim vrednostima. Ovaj metod objašnjava promene kroz tri komponente: (p, d, q).

Prva komponenta - p, predstavlja autoregresivnu komponentu modela AR(p). Ona predstavlja vezu (linearnu regresiju) između trenutne vrednosti vremenske serije i njenih prethodnih p-vrednosti, sa parametrima koji predstavljaju težine uticaja prethodnih vrednosti na trenutnu vrednost. AR(p) se može predstaviti jednačinom

GLAVA 2. OPIS KORIŠĆENIH METODA I METRIKA

2.1, gde je ϵ_t šum/greška (*eng.* white noise), ϕ su vrednosti AR parametara, y_t je vrednost serije u vremenu t, μ je prosek promenljive y, p je vrednost kašnjenja.

$$\begin{aligned}y_t &= \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t \\ \delta &= (1 - \phi_1 - \phi_2 - \cdots - \phi_p) \mu\end{aligned}\tag{2.1}$$

Druga komponenta - d, predstavlja operaciju diferenciranja ($I(d)$) primenjenu d puta na vremensku seriju, sa ciljem da serija postane stacionarna ¹. Diferenciranje (*eng.* differencing) predstavlja pojam promene između 2 uzastopne posmatrane vrednosti u originalnoj seriji. Ono se može predstaviti jednačinom 2.2.

$$y'_t = y_t - y_{t-1}\tag{2.2}$$

Treća komponenta - q, predstavlja komponentu klizajućih proseka (*eng.* moving averages) i služi da objasni trenutnu ϵ_t vrednost tj. grešku, u odnosu na prethodnih q vrednosti greške. MA(q) model se može predstaviti jednačinom 2.3 [15], gde θ vrednosti predstavljaju vrednosti MA parametara, q je vrednost kašnjenja.

$$\begin{aligned}\tilde{y}_t &= \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q} \\ \tilde{y}_t &= y_t - \mu\end{aligned}\tag{2.3}$$

Procesi mogu da imaju obe komponente: AR i MA. Kada se ukombinuju prethodne jednačine, finalni model je predstavljen jednačinom 2.4.

$$\tilde{y}_t - \phi_1 \tilde{y}_{t-1} - \cdots - \phi_p \tilde{y}_{t-p} = \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q}\tag{2.4}$$

ARIMA može da ima i sezonske komponente: (P, D, Q), kada se nazima SARIMA. Sezonske komponente predstavljaju regularne obrazce promena koje se ponavljaju svakih S vremenskih perioda, gde je S broj vremenskih perioda dok se obrazac opet ne ponovi. Detaljnije o SARIMA modelu se može naći u [6].

2.2 Prophet

Prophet je alat za predviđanje ciljne promenljive kod vremenskih serija. Razvijen je od strane Facebook zajednice [12] i može da generiše prognoze razumnog kvaliteta. Alat je otvorenog tipa (*eng.* open-source) i zasnovan je na aditivnom modelu, gde

¹Vremenska serija je stacionarna kada sve njene vrednosti fituju istoj raspodeli verovatnoće, nevezano za vreme [15]. Odnosno, kada serija nema trend ili sezonske efekte i kad su statistička svojstva konstanta kroz vreme [11].

GLAVA 2. OPIS KORIŠĆENIH METODA I METRIKA

se nelinearni trendovi fituju sa godišnjom, nedeljnom i dnevnom sezonalnošću (*eng. seasonality*). Ono što Prophet ima kao opciju je uključivanje efekta praznika u model. Robustan je na nedostajuće vrednosti i promene u trendu, i tipično dobro obrađuje vrednosti van granica (*eng. outliers*) [17].

Originalni rad u kome autori predstavljaju Prophet [13], pomenuti model je predstavljen jednačinom 2.5.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (2.5)$$

U jednačini 2.5, $g(t)$ predstavlja funkciju trenda koja modeluje neperiodične promene vrednosti ciljne promenljive, $s(t)$ predstavlja periodične promene (npr. sezonske promene dnevnog ili godišnjeg nivoa), $h(t)$ predstavlja efekat praznika koji mogu da se javljaju u neregularnim trenucima i da traju 1 ili više dana, ϵ_t se odnosi na grešku i obuhvata sve promene neuhvaćene modelom. Pretpostavka koju autori pominju u radu je da ϵ_t ima normalnu raspodelu. Pored navedenih opcija, moguće je kreirati sopstveni sezonski efekat (npr. mesečni) koji se generiše koristeći parcialne Furijeove sume, kao i dosta drugih manuelnih opcija². Prophet ima podršku za programski jezik Python i programski jezik R.

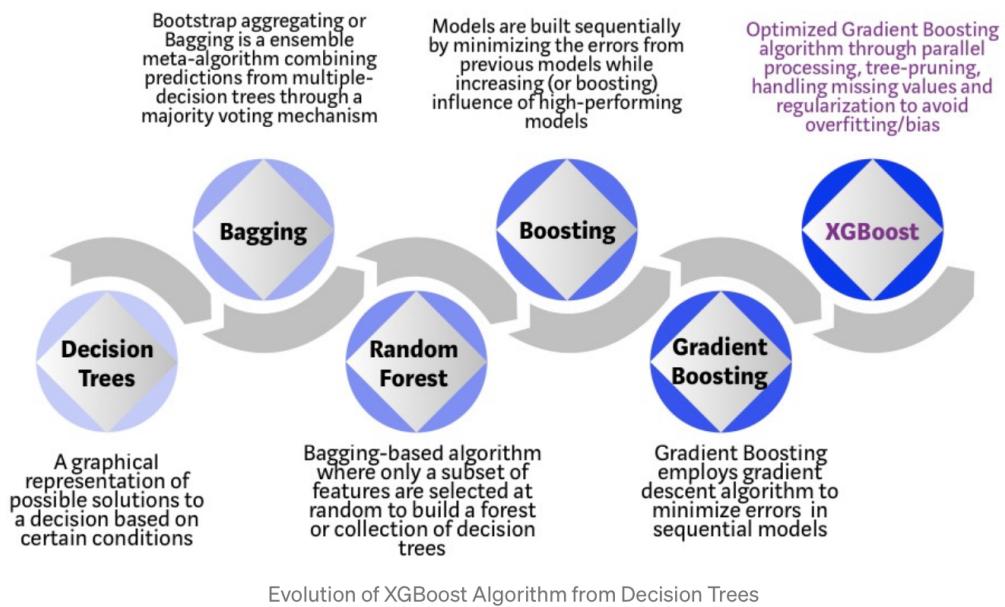
2.3 XGBoost

XGBoost (*eng. Extreme Gradient Boosting*) je metod ekstremnog gradijentnog pojačavanja. Osnovna ideja metoda pojačavanja je da se ansambl izgradi dodavanjem jednog po jednog modela, gde svaki model uči tako da što bolje nadomesti mane trenutnog skupa modela. AdaBoost je prvi algoritam pojačavanja i radi po principu uvećavanja težina onih instanci koje su pogrešno klasifikovane, pa se sledeći model fokusira na te instance kako bi se celokupno stanje poboljšalo.

Osim pojačavanja, postoje i gradijentna pojačavanja i ona su po pitanju preciznosti među najboljim metodama mašinskog učenja. Osnovna ideja ovih metoda dolazi iz gradijentnih optimizacionih problema, gde se tekuće rešenje popravlja dodavanjem vektora proporcionalnog negativnoj vrednosti gradijenta funkcije koja se minimizuje [9].

Ekstremno gradijentno pojačavanje predstavlja ansambl sistem pojačavanja zasnovan na drvetima odlučivanja. Evolucija algoritama je predstavljena na slici 2.1 i može se videti da je regularizacija jedna od novina kod XGBoost metode.

²Više se može pročitati na <https://facebook.github.io/prophet/>



Slika 2.1: Razvoj algoritama. Izvor: <https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d>

Sam algoritam je poznat po svojoj skalabilnosti, dobrom ponašanju sa retkim (eng. sparse) podacima, distribuiranosti i algoritam je podržan od strane nekoliko programskih jezika, među kojima je i Python [1].

2.4 Metrike

Kroz rad je ispráčeno nekoliko metrika za evaluaciju modela. Korišćene metrike su:

- Srednja apsolutna greška (MAE)
- Srednje kvadratna greška (MSE)
- Koren srednje kvadratne greške (RMSE)
- Simetrična srednja apsolutna procentualna greška (SMAPE)

Ključna metrika koja je praćena je SMAPE, ali i ostale metrike su propraćene kroz modele. U nastavku će biti detaljnije opisana SMAPE metrika.

SMAPE

SMAPE (*eng.* Symmetric Mean Average Percentage Error) metrika je izmenjena za potrebe projekta i formula po kojoj je računata je predstavljena jednačinom 2.6.

$$SMAPE = \frac{\sum_{t=1}^N |F_t - A_t|}{\sum_{t=1}^N \max(A_t, F_t)} \quad (2.6)$$

U formuli F_t predstavlja prediktovanu vrednost u vremenu t, A_t predstavlja stvarnu vrednost u vremenu t, a N je broj instanci. Manje vrednosti ove metrike su pogodnije za predikcije i u praksi se pokazalo da modeli sa vrednostima manjim od 0.3 imaju nekog smisla.

Osim metrika za evaluaciju modela, korišćena je Akaike Information Criterion³ (AIC) metrika za evaluaciju ARIMA modela kod vremenskih serija. Ona je računata ugrađenom funkcijom u programskom jeziku Python. Manja vrednost ove metrike označava bolji model.

2.5 Impact Encoding

Impact Encoding predstavlja jednu vrstu enkodiranja kategoričkih atributa u redne vrednosti. Ono predstavlja nadogradnju na Target Encoding [10] i praktičnije je od one-hot enkodiranja zbog toga što dodaje samo 1 kolonu koja predstavlja izlaz, za razliku od one-hot enkodiranja. Sastoji se od 3 koraka: Target Encoding, regularizacija i normalizacija.

Target Encoding je deo koji svaku vrednost kategoričkog atributa zamjenjuje prosečnom vrednošću koju baš ta vrednost kategoričkog atributa ima u koloni ciljne promenljive. Recimo, ako je kategorički atribut boja i ona može da uzme vrednosti: crvena i plava, onda crvenu enkodiramo sa prosekom vrednosti koju crvena ima u koloni ciljne promenljive. Ovo se može zapisati sledećom formulom u kojoj T predstavlja vrednost kategoričkog atributa c_v u koloni ciljne promenljive:

$$c_v \rightarrow \frac{\sum T_{cv}}{|c_v|}$$

³https://en.wikipedia.org/wiki/Akaike_information_criterion

GLAVA 2. OPIS KORIŠĆENIH METODA I METRIKA

Deo regularizacije enkodiranja se bavi problemom da ne verujemo baš svakom mapiranju vrednosti jednako. Ovaj problem se javlja u praksi kada vrednost ciljne promenljive za jednu vrednost kategoričkog atributa može da bude ogromna, a da te vrednosti atributa ima jako malo; naspram puno vrednosti kategoričkog atributa čije su vrednosti koje odgovaraju ciljnoj promenljivoj dosta manje. Želimo da izbalansiramo ovu situaciju i to se rešava uvođenjem parametra $\alpha := \max |c_v|$, koji predstavlja maksimum broja pojavljivanja vrednosti kategoričkog atributa. Tada enkodiranje postanje mapiranje atributa u:

$$c_v \rightarrow \frac{|c_v|\mu_{c_v} + \alpha\mu_c}{|c_v| + \alpha}$$

Deo normalizacije rešava problem kada Impact Encoding radimo nad trening podacima, a u test skup nam dođe vrednost atributa koja nije viđena u trening skupu dok smo vršili mapiranje vrednosti. Rešenje u ovoj situaciji je normiranje sa prosekom svih klasa:

$$c_v \rightarrow \frac{\frac{|c_v|\mu_{c_v} + \alpha\mu_c}{|c_v| + \alpha}}{\mu_c}$$

Ovako nepoznatim instancama može da se dodeli vrednost 1, koja implicira da je impact očekivana srednja vrednost svih klasa i da nema odstupanja od populacije.

Glava 3

Prikaz rada metoda i rezultati

3.1 Podaci

Podaci korišćeni u radu su prikupljeni od jedne Švedske kompanije. Predstavljaju realne podatke iz industrije, a važno je napomenuti da su dobijeni sa samo jednog izvora koji je dostupan njihovim korisnicima. Radi se o novom načinu rezervisanja popravki automobila na različitim lokacijama, putem interneta (*eng.* booking portal). Portal je krenuo sa radom krajem 2019. godine, tako da je količina podataka ograničena. Takođe, 2020. godina je bila udarna godina virusa Covid19 [7], tako da ograničena količina dostupnih podataka i istraživanje nad njima možda bude od posebnog značaja u budućnosti, pre svega kada pomenuti booking portal bude izvor veće količine informacija.

Podaci se sastoje od informacija kao što su: koja marka automobila je došla u koju registrovanu automehaničarsku radnju, kog datuma i od nekih informacija vezanih za popravku. Svi korišćeni kategorički atributi su morali biti zamaskirani i kodirani kombinacijama slova zbog privatnosti firme, a svi numerički atributi su normalirani. Na slici 1.2 se može videti primer korišćenog skupa podataka za predviđanje promenljive *demand_value*, na granularnosti marki automobila i automehaničarskih radnji, na nivou nedelja. Atributi *x_units* i *x_unit_cost* su vezani za informacije o konkretnim rezervacijama, od kojih u modelima nije korišćen *x_units*, kako ne bi dolazilo do potencijalnog curenja informacija u modelu.

Obogaćivanje podataka eksternim podacima

Atributi *number_of_competitors* i *reachable_population* su generisani na osnovu firminih informacija o lokacijama automehaničarskih radnji. Za svaku radnju su poznate geografska širina i geografska dužina. Atribut *reachable_population* je kreiran od strane zaposlenih u firmi, na osnovu javnih statističkih podataka o populaciji stanovništva u regionima radnji. On predstavlja sumu populacije velikih regionala u radijusu koji predstavlja distancu pređenu za 40 minuta vožnje, brzinom 60 km/h, od lokacije radnje.

Atribut *number_of_competitors* je dodat u sklopu ovog rada. On predstavlja broj drugih radnji (konkurenata) u radijusu od 80km od automehaničarske radnje. Vrednost od 80km je uzeta kao neka racionalna vrednost za voženje do neke recimo jeftinije radnje ili do radnje koja nije u drugom gradu. Vrednost ovog atributa je kreirana korišćenjem Open Street Map (OSM)¹, odnosno besplatne mape sveta. OSM ima besplatan servis za dohvatanje raznih informacija na osnovu zadatah parametara, korišćenjem njihovog upitnog jezika. Servis se zove Overpass API² i zadavanjem lokacija radnji i specijalnog taga '*shop*'='car_repair' dobijene su informacije o radnjama u okolini u *json* formatu. Primer koda korišćen za dohvatanje informacija o radnjama je prikazan u listingu 3.1.

```
1 overpass_url = "http://overpass-api.de/api/interpreter"
2
3 def overpass_query(latitude, longitude, km_around):
4     overpass_query = "[out:json];(node(around:{}{},{});['shop']='car_repair'
5     ']);;out;".format(km_around, latitude, longitude)
6
7     while True:
8         try:
9             response = requests.get(overpass_url, params={'data':
10             overpass_query})
11             garages_around = response.json()
12             return garages_around
13         except:
14             print("Try again...")
```

Kod 3.1: Funkcija za dohvatanje informacija o drugim radnjama u okolini.

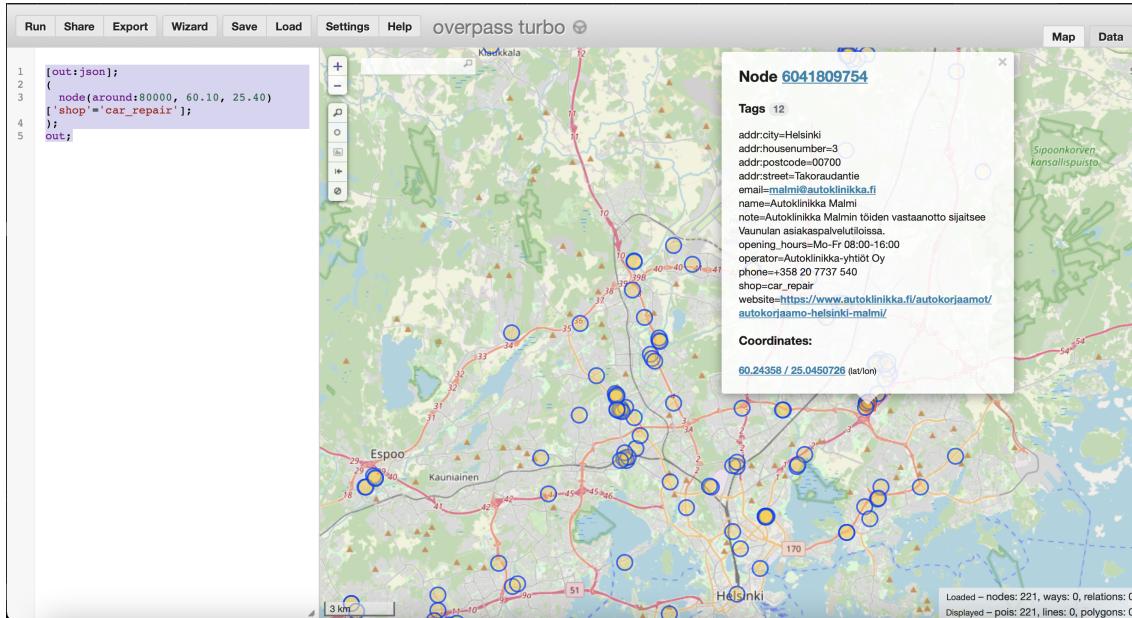
Primer informacija koje se dobiju upitom ka Overpass API prikazan je na slici 3.1. Za atribut skupa podataka u ovom radu, korišćen je samo ukupan broj radnji u

¹<https://www.openstreetmap.org>

²<https://overpass-turbo.eu>

GLAVA 3. PRIKAZ RADA METODA I REZULTATI

okolini.



Slika 3.1: Overpass API primer upita.

Primer podataka koji je korišćen za vremenske serije na dnevnom nivou predstavljen je na slici 3.2, a na nedeljnom nivou na slici 3.3. Bitno je napomenuti da su datumi prebačeni u nedelje u godini, tako da se vodilo računa o tome da kalendarski dan bude u tekućoj kalendarskoj godini, npr. da 30. decembar 2019. godine bude u poslednjoj nedelji 2019. godine, a ne u prvoj nedelji 2020. godine. Odluka da tako bude je napravljena u dogovoru sa firmom, da bude usaglašena sa kalendarskom godinom i kalkulacijama na nivou godine. Primer koda koji sređuje prebacivanje, kao i izvor sa svim podacima korišćenim u radu, dat je u Github repozitorijumu projekta, na adresi: <https://github.com/mandja96/matf-master-rad>.

3.2 Dnevni nivo - na nivou države

U ovom poglavlju biće predstavljeni propraćeni koraci i dobijeni rezultati nad dnevnim podacima, koji predstavljaju univariantnu vremensku seriju. Podaci su agregirani po datumima, za nivo cele države. Biće izložena tri pristupa: Arima, Prophet i XGBoost, koji su opisani u poglavlju 2. Razlog posmatranja vremenske serije nad agregiranim podacima je pre svega da se isproba ponašanje metoda, ali

GLAVA 3. PRIKAZ RADA METODA I REZULTATI

| demand_value | date |
|---------------------|------------|
| 0.09803921568627451 | 2019-12-18 |
| 0.24509803921568626 | 2019-12-19 |
| 0.4019607843137255 | 2019-12-20 |
| 0.49019607843137253 | 2019-12-23 |
| 0.5196078431372549 | 2019-12-27 |
| 0.0196078431372549 | 2019-12-28 |
| 0.5 | 2019-12-30 |

Slika 3.2: Primer korišćenih podataka za dnevnu vremensku seriju.

| year_week | demand_value |
|-----------|---------------------|
| 201951 | 0.14366729678638943 |
| 201952 | 0.2948960302457467 |
| 202001 | 0.13988657844990549 |
| 202002 | 0.332703213610586 |
| 202003 | 0.7901701323251418 |
| 202004 | 0.9206049149338374 |
| 202005 | 0.9149338374291115 |

Slika 3.3: Primer korišćenih podataka za nedeljnu vremensku seriju.

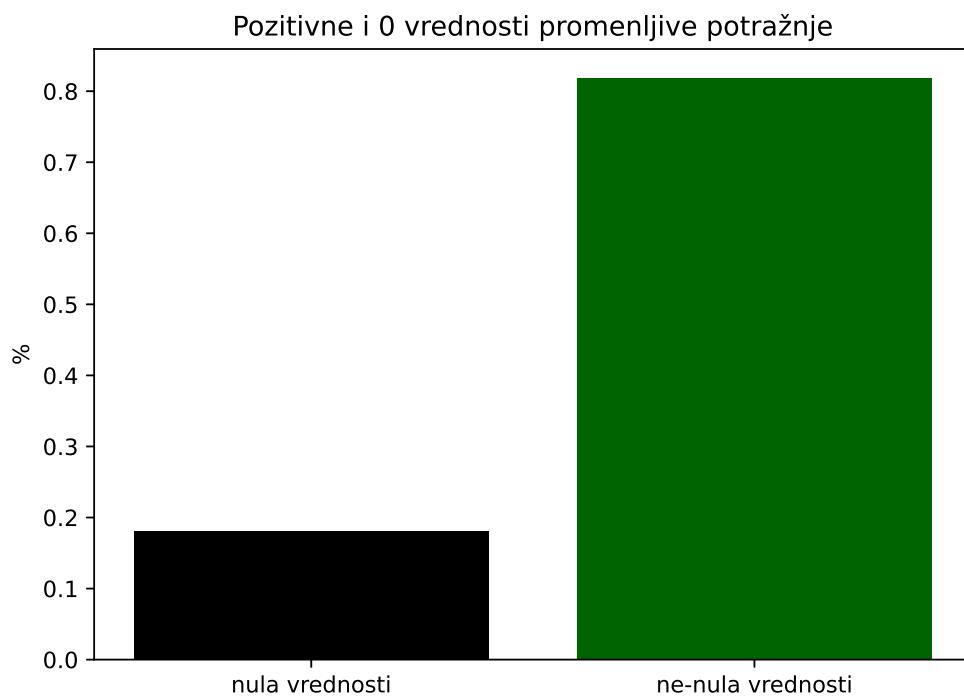
i da se upoznaju podaci i generalni nedostatak podataka za vremensku seriju nad zasebnim radnjama i markama (većim granularnostima-manjoj agregaciji).

Preprocesiranje dnevnih podataka

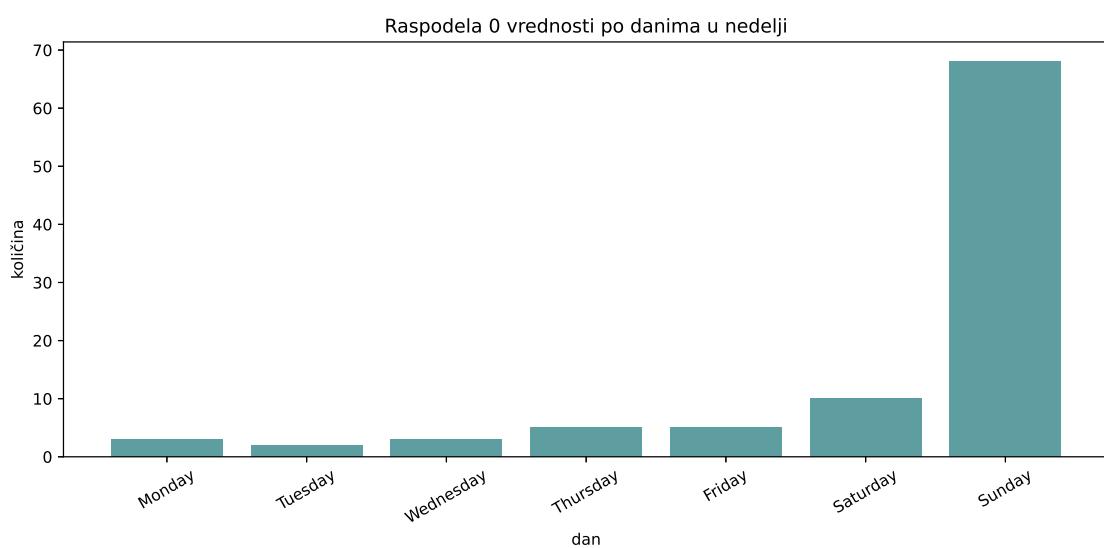
Pre svega je ispitano koliko datuma nedostaje u vremenskoj seriji u vremenskom periodu 18. decembar 2019 - 29. maj 2021. Broj nedostajućih dana je 96 i ti dani su popunjeni vrednošću 0. Odnos nula i ne-nula dana prikazan je na histogramu 3.4, a raspodela nedostajućih dana po danima nedelje je prikazana na histogramu 3.5.

Kao što se može videti sa histograma, većina 0 vrednosti pripada nedeljama. Ostatak datuma koji imaju vrednost nula su pretežno neki praznici koji su neradni dani (oko Nove godine, Uskrsa, državnog praznika...). Zbog velike zastupljenosti

GLAVA 3. PRIKAZ RADA METODA I REZULTATI



Slika 3.4: Procenat pozitivnih i nula vrednosti promenljive potražnje, nakon popunjavanja nedostajućih datuma.



Slika 3.5: Raspodela nula vrednosti promenljive potražnje po danima u nedelji.

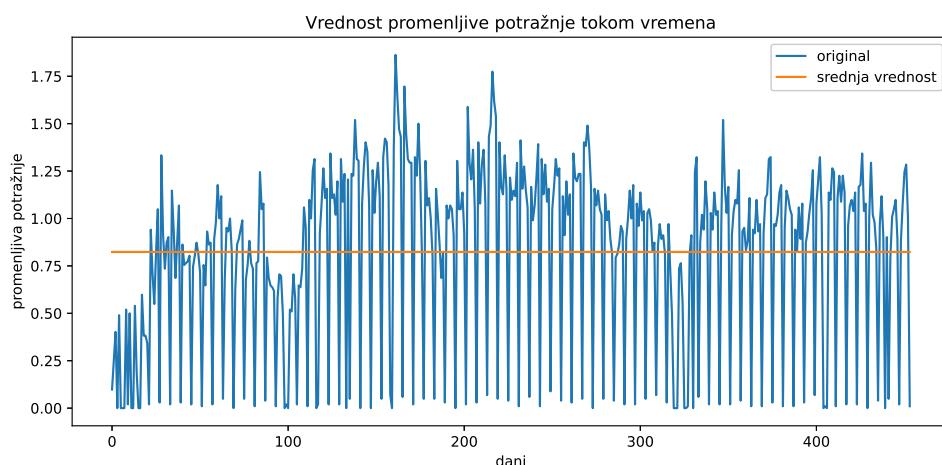
GLAVA 3. PRIKAZ RADA METODA I REZULTATI

0 vrednosti kod nedelje kao dana u sedmici, sve nedelje su izbačene iz vremenske serije i serija je posmatrana kao da postoji 6 dana u sedmici. Veličina posmatranih dana je nakon toga postala 454.

Arima

Kod Arime je bilo bitno ispitati neka svojstva vremenske serije kao što su nedostajuće vrednosti (datumi), stacionarnost, sezonski efekti. Takođe, bilo je bitno ispitati ACF i PACF plotove, kako bi se odredili parametri (p, d, q) koji konstruišu model, ali ispitati i kako se ponaša sezonska Arima (SARIMA) nad podacima.

Nakon pomenutog preprocesiranja podataka, vremenska serija ima izgled prikazan na grafiku 3.6.



Slika 3.6: Vremenska serija bez nedelja.

Prva sledeća stvar je bila ispitivanje stacionarnosti serije, jer je na osnovu toga doneta odluka da li seriju treba diferencirati ili ne. U ove svrhe korišćen je statistički **Augmented Dickey-Fuller** test. Ovaj test se može pronaći implementiran u paketu *statsmodels* pod nazivom *adfuller*. Vrednost testa nad dnevnim podacima bez nedelja je vratila vrednosti:

| | |
|-----------------------------|------------|
| Test Statistics | -3.674781 |
| p-value | 0.004482 |
| No. of lags used | 18.000000 |
| Number of observations used | 435.000000 |

GLAVA 3. PRIKAZ RADA METODA I REZULTATI

| | |
|----------------------|-----------|
| critical value (1%) | -3.445473 |
| critical value (5%) | -2.868207 |
| critical value (10%) | -2.570321 |

Odavde se može zaključiti da je p-vrednost manja od 0.05 i da se nulta hipoteza može odbaciti, što nam govori da je serija stacionarna i da nema potrebe za differenciranjem. Dakle, parametar $d=0$. Takođe, radi dodatne provere, iz paketa **pmdarima** iskorišćena je ugrađena funkcija *ndiffs*, koja vraća vrednost parametra d (u ovom slučaju 0).

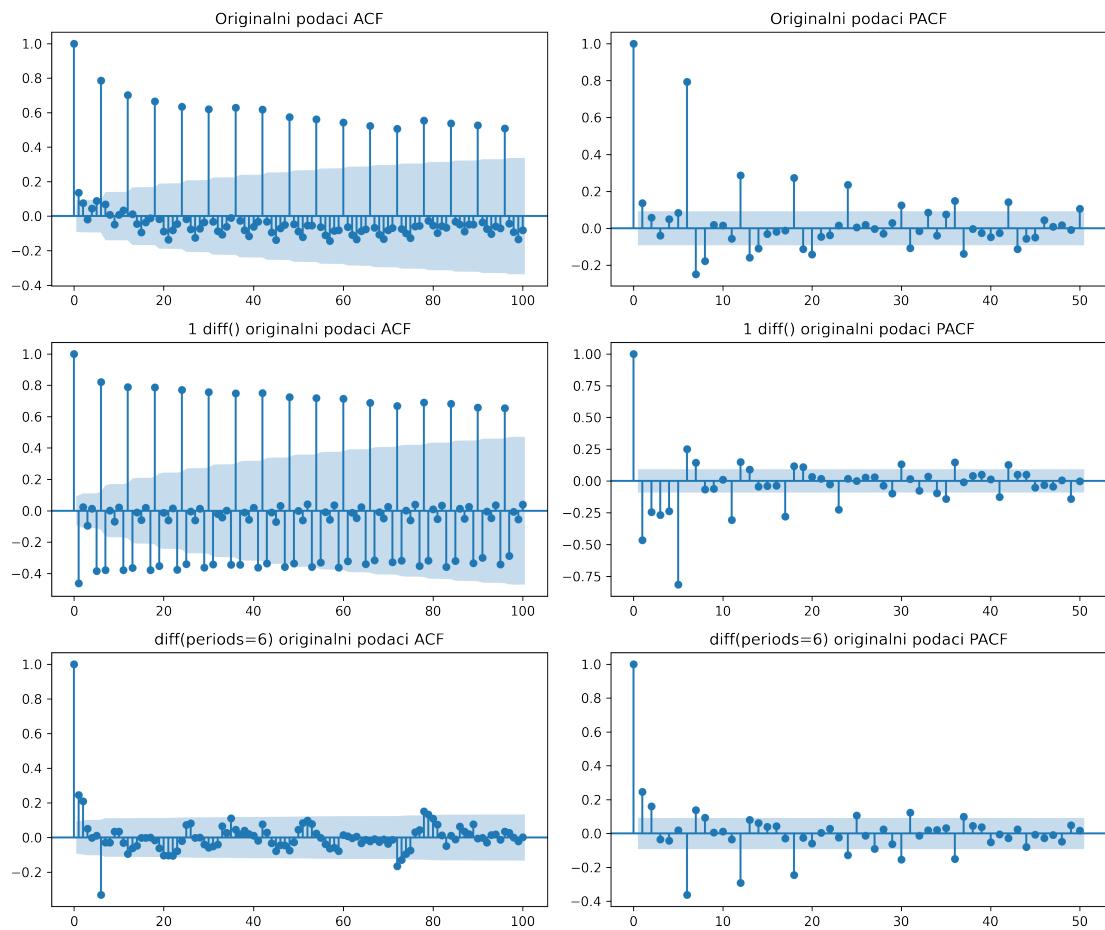
Sledeće je potrebno ispitati koje vrednosti parametri p i q mogu potencijalno da imaju. Za ispitivanje ovih parametara nam mogu pomoći ACF i PACF plotovi. Na grafiku 3.7 se mogu videti plotovi za ove podatke. Primećujemo da originalni podaci imaju pikove na svakih 6 stubića. To nam govori da postoji možda neka sezonalnost, tako da su propraćeni i ACF/PACF plotovi sa razlikom trenutne vrednosti i vrednosti u trenutku 6 vrednosti u prošlosti. Za dobijanje ovih grafika korišćene su funkcije **plot_acf** i **plot_pacf** iz paketa *statsmodels*.

Na osnovu knjige [6], sa plotova se može zaključiti da se radi o ARIMA(0, 0, 0)(0, 0, 1)[6] modelu. To je sezonska Arima (SARIMA) koja ima MA=1, ali sa kašljenjem od 6 dana. Dakle, ovo sugerire da za predviđanje ponедeljka treba gledati u vrednosti prethodnog ponedeljka, za predviđanje utorka vrednosti prethodnog utorka itd, jer je na vrednosti 6 kod ACF plota najviši stubić.

Pokretanjem **auto_arima** funkcije iz paketa *pmdarima*, najbolji model je ARIMA(2, 0, 0)(2, 0, 1)[6]. On je dobijen gledanjem u minimalnu vrednost AIC metrike opisane u poglavlju 2. Kako su dva SARIMA modela sugerisala da treba gledati u jednu ili 2 vrednosti 6 dana u prošlosti, ispitani je i model ARIMA(6, 0, 6) i model ARIMA(9, 0, 6) koji je dobijen **auto_arima** funkcijom kada se isključi sezonski parametar. Pored njih, iz radoznalosti, ispitani su i modeli ARIMA(1, 0, 1), ARIMA(1, 0, 0), ARIMA(1, 0, 1)(1, 0, [1,2])[6]. Vrednosti AIC metrike su prikazane u tabeli 3.1, vrednosti metrika evaluacije u tabeli 3.2, a autokorelacioni grafik reziduala najboljeg modela na slici 3.8 (dobijen je funkcijom *plot_diagnostics* nad modelom ARIMA(1, 0, 1)(1, 0, [1, 2])[6]). Reziduali treba da nemaju korelaciju, kako bi predstavljali samo šum (*eng. white noise*).

Detaljan način obučavanja ARIME se može pronaći u kodovima na github repozitorijumu projekta. Princip obučavanja je bio da se u trening skupu na početku nalaze svi dani pre 1. marta 2021. godine, a da se zatim predviđa 1 po 1 dan u budućnost, nakon čega se u trening skup dodaje 1 dan. Modeli su evaluirani na

GLAVA 3. PRIKAZ RADA METODA I REZULTATI



Slika 3.7: ACF i PACF grafici. Na x-osi su lagovi, a na y-osi vrednost korelacije.

Tabela 3.1: SARIMA/ARIMA vrednosti AIC metrike

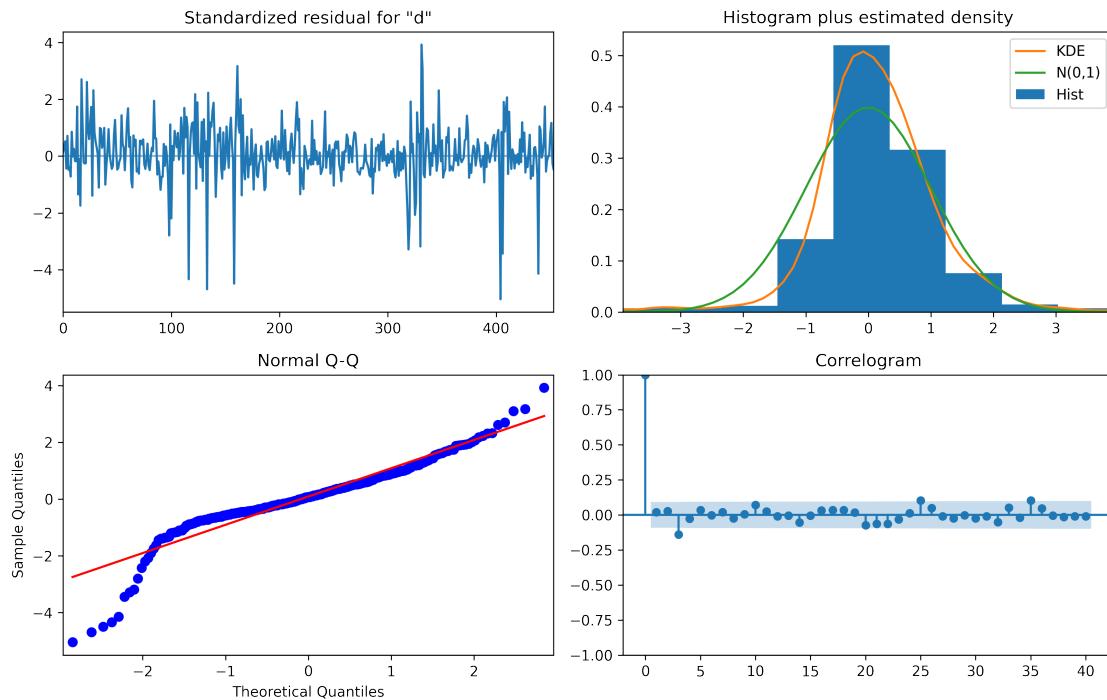
| model | AIC |
|--------------------------------|--------|
| ARIMA(1, 0, 1) | 567.39 |
| ARIMA(1, 0, 0) | 598.30 |
| ARIMA(6, 0, 6) | 14.31 |
| ARIMA(9, 0, 6) | 14.63 |
| ARIMA(1, 0, 1)(1, 0, [1,2])[6] | 3.81 |
| ARIMA(2, 0, 0)(2, 0, 1)[6] | 8.57 |

osnovu predviđanja jednog dana u napred od 1. marta 2021. godine, treniranjem nad svim podacima pre. Takođe, predikcije su vršene i po nekoliko dana unapred odjednom (konkretno 28 dana), da se vidi ponašanje modela. Na grafiku 3.9 se može videti ponašanje dva modela.

GLAVA 3. PRIKAZ RADA METODA I REZULTATI

Tabela 3.2: SARIMA/ARIMA metrike evaluacije

| model | mae | mse | rmse | smape |
|--------------------------------|----------|----------|----------|----------|
| ARIMA(1, 0, 1) | 0.389531 | 0.230040 | 0.479625 | 0.379000 |
| ARIMA(1, 0, 0) | 0.382899 | 0.215804 | 0.464547 | 0.375804 |
| ARIMA(6, 0, 6) | 0.563536 | 0.094383 | 0.307218 | 0.213063 |
| ARIMA(9, 0, 6) | 0.209452 | 0.095371 | 0.308821 | 0.224278 |
| ARIMA(1, 0, 1)(1, 0, [1,2])[6] | 0.192394 | 0.094991 | 0.308206 | 0.208225 |
| ARIMA(2, 0, 0)(2, 0, 1)[6] | 0.200078 | 0.092266 | 0.303753 | 0.215694 |



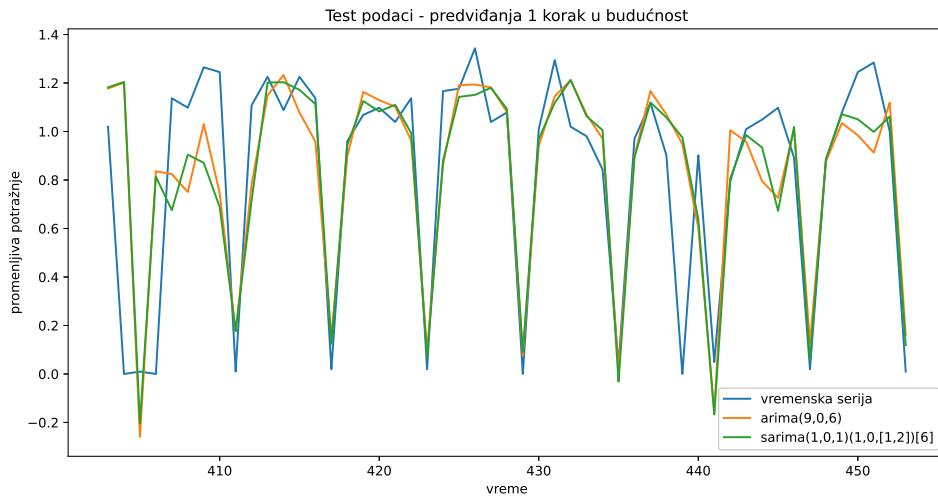
Slika 3.8: Provera normalnosti i odsustva korelacije među rezidualima modela.

Prophet

Nad istim podacima i na isti način, isprobao je i Facebook-ov alat Prophet. Puštena su 4 modela, sa uključenim praznicima u Švedskoj, pošto je to omogućena opcija kod ovog metoda.

Metrike za evaluaciju modela su prikazane u tabeli 3.3, a grafik predviđanja na slici 3.10. Sa grafika se može videti kako postoji jedan značajniji pad. Razlog tome je uračunavanje efekta praznika u modele, što je potvrdilo ideju da potencijalno treba uzimati u obzir praznike i u drugim modelima, ako je to moguće.

U tabeli su označke koje predstavljaju informaciju da li su u model uključeni



Slika 3.9: Ponašanje 1 po 1 predviđanja dva modela kod arima metode.

Tabela 3.3: Prophet metrike evaluacije

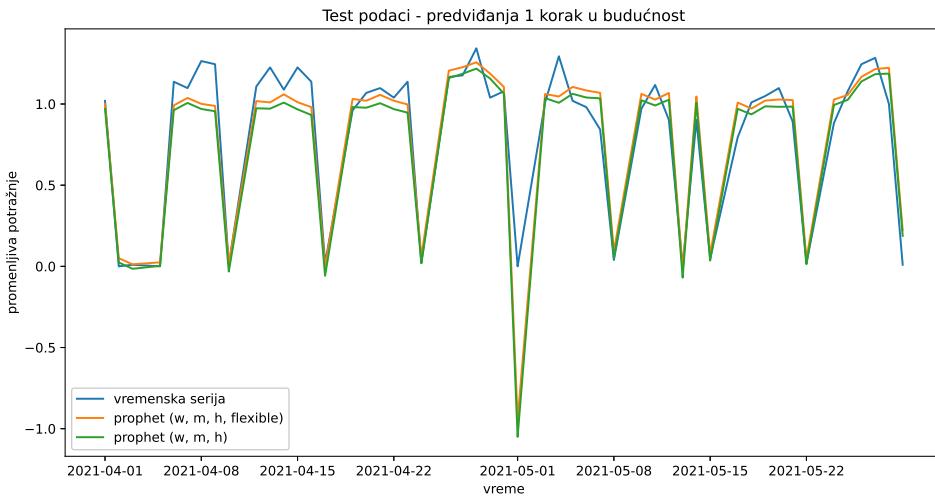
| model | mae | mse | rmse | smape |
|--------------------------------|----------|----------|----------|----------|
| prophet (w, m, h) | 0.120943 | 0.038284 | 0.195662 | 0.141016 |
| prophet (w, m, h, flexible) | 0.113431 | 0.032307 | 0.179741 | 0.129801 |
| prophet (d, w, m, h) | 0.119556 | 0.035445 | 0.188267 | 0.139296 |
| prophet (d, w, m, h, flexible) | 0.113918 | 0.032443 | 0.180120 | 0.130383 |

sezonski efekti i praznici. Oznake su: w - uključen nedeljni sezonski efekat, m - uključen mesečni sezonski efekat, h - uključeni praznici, d - uključen dnevni sezonski efekat i flexible označava postavljanje parametra *changepoint_prior_scale*=0.09, što znači postavljanje trenda da bude više fleksibilan ³.

XGBoost

XGBoost je isprobao nad istim podacima, ali je sa njim urađen jedan eksperiment. Atributi za XGBoost nad dnevnim podacima su zapravo vrednosti promenljive potražnje (*eng. demand_value*) iz prethodnih nekoliko dana u prošlosti. Testirano je kako XGBoost radi sa 30 vrednosti iz prošlosti, sa tri dodata atributa koja predstavljaju praznike u Švedskoj. Praznici u Švedskoj su uzeti iz paketa *holidays*, i 3 atributa predstavljaju:

³https://facebook.github.io/prophet/docs/trend_changepoints.html



Slika 3.10: Ponašanje 1 po 1 predviđanja dva modela kod prophet metode. Razlog zašto test podaci počinju tek 1. aprila, a ne 1. marta, je zbog toga što je testirano predviđanje do 28 dana u napred, tako da konkateniranje skupova izbacuje nedostajuće vrednosti koje postoje 28 dana nakon 1. marta.

Tabela 3.4: XGBoost metrike evaluacije

| mae | mse | rmse | smape |
|----------|----------|----------|----------|
| 0.112533 | 0.033992 | 0.184370 | 0.127315 |

- *number_of_holidays_next_day* - broj praznika sledećeg dana
- *number_of_holidays_previous_day* - broj praznika prethodnog dana
- *number_of_holidays* - broj praznika u tekućem dana

Validacioni set za XGBoost je predstavljao 20% test podataka koji kreću od 1. marta 2021. godine i korišćen je parametar *early_stopping_rounds*=30, kako bi se sprečilo preprilagođavanje. Metrike evaluacije za ovako dobijeni model su predstavljene u tabeli 3.4. Važnost atributa (*eng. feature importance*) ovakvog modela je predstavljena ispod:

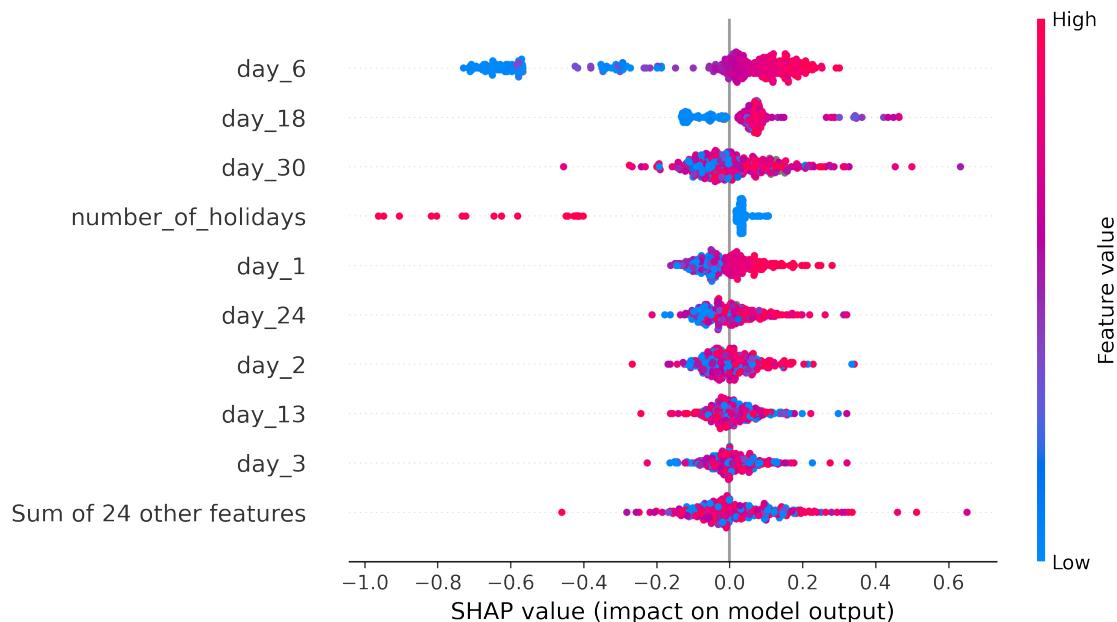
```
[0.00794075, 0.00629401, 0.00511439, 0.00234501, 0.00245581,
 0.42707652, 0.0020948 , 0.00207992, 0.00157057, 0.0035512 ,
 0.00304402, 0.00182108, 0.00567753, 0.00143691, 0.00109723,
 0.00155572, 0.0033229 , 0.08218686, 0.00293316, 0.00355283,
 0.00354679, 0.00391645, 0.00104043, 0.01375661, 0.0011694 ,
```

GLAVA 3. PRIKAZ RADA METODA I REZULTATI

0.00137136, 0.00357201, 0.00281513, 0.00125128, 0.00372967,
0.39017874, 0.0065008 , 0.]

Odavde se može zaključiti da XGBoost najveću vrednost daje vrednosti ciljne promenljive iz prethodnih 6 dana, iz prethodnih 18 dana i vrednosti broja praznika tekućeg dana.

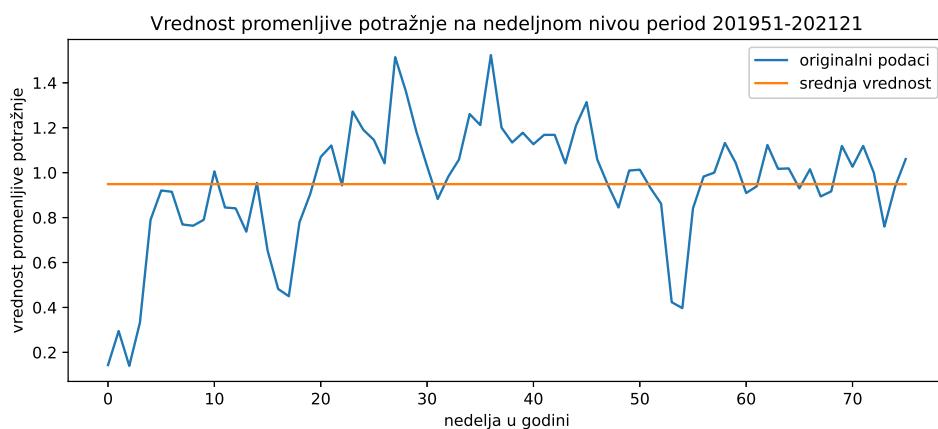
Pored ovog eksperimenta, urađena su još 2 eksperimenta. Prvi je bio dodavanje na postojeće atrbute još jedan atrbut koji predstavlja prosek vrednosti prethodnih 6 dana. Drugi eksperiment je na prvobitne atrbute dodao prosek razlike od srednje vrednosti za prethodnih 6 dana (pokušavajući da se imitira model arime sa MA=6 koeficijentom). Ovi eksperimenti su bili bezuspešni, sa atrbutima koji nisu doprineli poboljšanju modela. Na SHAP grafiku 3.11 se može videti kako 6. dan iz prošlosti doprinosi smanjenju vrednosti, a broj praznika u danu povećanju vrednosti ciljne promenljive. Ova analiza je urađena samo na celom trening setu osnovnog modela za koji su izložene metrike, kako bi se ispitalo ponašanje.



Slika 3.11: SHAP vrednosti nad trening skupom.

3.3 Nedeljni nivo - na nivou države

Primer podataka agregiranih na nedeljnog nivou, može se videti na slici 3.3. Ta vremenska serija je predstavljena na grafiku 3.12. Ukupan broj dostupnih nedelja je 76. Prvo što se može primetiti kod ove vremenske serije su 2 veća pada vrednosti. Jedan se desio oko Uskrsa 2020. godine, a drugi se desio oko Nove godine.



Slika 3.12: Nedeljna vremenska serija.

Arima i Prophet

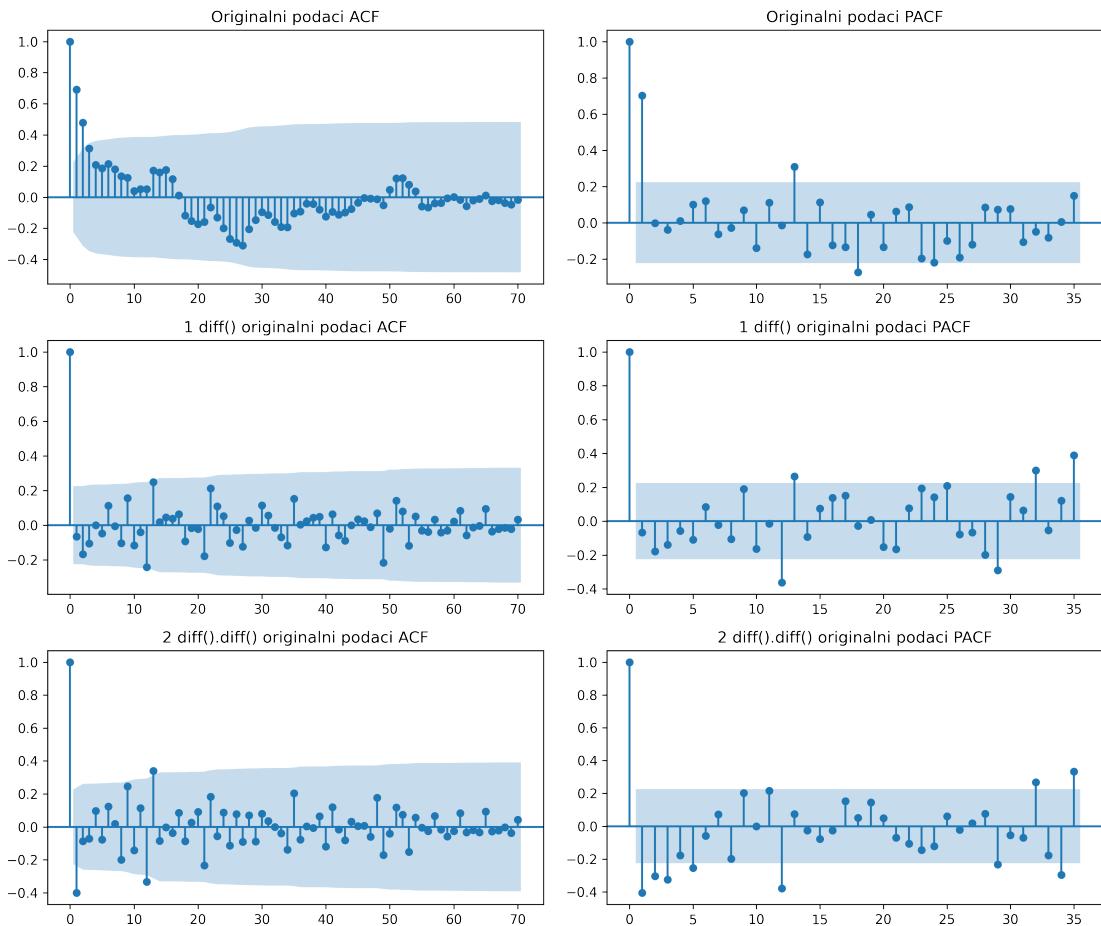
Kao i kod primera nad dnevnim podacima, prva ispitivana stvar je stacionarnost vremenske serije. Statistički **Augmented Dickey-Fuller** test je korišćen i vrednosti rezultata nad nedeljnim podacima su sledeći:

| | |
|-----------------------------|-----------|
| Test Statistics | -3.915460 |
| p-value | 0.001924 |
| No. of lags used | 0.000000 |
| Number of observations used | 75.000000 |
| critical value (1%) | -3.520713 |
| critical value (5%) | -2.900925 |
| critical value (10%) | -2.587781 |

Može se zaključiti da je serija stacionarna i da bi vrednost parametra d trebala da bude 0, tj. da nema potrebe za diferenciranjem podataka. Kako se radi o nedeljnim podacima, za svaku nedelju neki podaci su sigurno dostupni, tako da ne postoje

GLAVA 3. PRIKAZ RADA METODA I REZULTATI

nedostajuće nedelje. Na grafiku 3.13 su prikazani grafici ACF i PACF, sa ciljem da se bolje razume ponašanje serije i da se odredе najpogodnije vrednosti parametara p i q . Sa grafika vremenske serije se može zaključiti da nema nekih lako uočljivih sezonskih efekata, za dostupnu količinu podataka, a to potvrđuju i ACF/PACF plotovi. Ono što se praćenjem literature može zaključiti je da se radi o modelu



Slika 3.13: ACF i PACF grafici. Na x-osi su lagovi, a na y-osi vrednost korelacije.

ARIMA(1, 0, 2) najverovatnije. Stubića izvan značajne zone na PACF plotu, do prvog sledećeg stubića unutar zone, ima 1 što određuje AR parametar. Na grafiku ACF se nalaze 2 stubića izvan značajne zone, tako da nam je to neki indikator da je parametar MA=2. Model koji je funkcija **auto_arima** vratila kao najbolji prema AIC metrići je ARIMA(1, 0, 0), dakle čist autoregresivni model koji predikcije pravi isključivo na jednoj prethodnoj vrednosti cilje promenljive. Pored ovih modela isprobani modeli su i: ARIMA(1, 1, 1), ARIMA(12, 1, 0) i ARIMA(13, 0, 2). Za model ARIMA(13, 0, 2) razlog je što PACF pokazuje značajan stubić na vrednosti

GLAVA 3. PRIKAZ RADA METODA I REZULTATI

Tabela 3.5: AIC metrika za ARIMA modele nad nedeljnim podacima.

| model | AIC |
|-----------------|---------------------|
| ARIMA(1, 0, 0) | -45.56447485932554 |
| ARIMA(1, 1, 1) | -43.27962567290584 |
| ARIMA(1, 0, 2) | -41.93537038126968 |
| ARIMA(12, 1, 0) | -39.918282940344085 |
| ARIMA(13, 0, 2) | -35.86521951018952 |

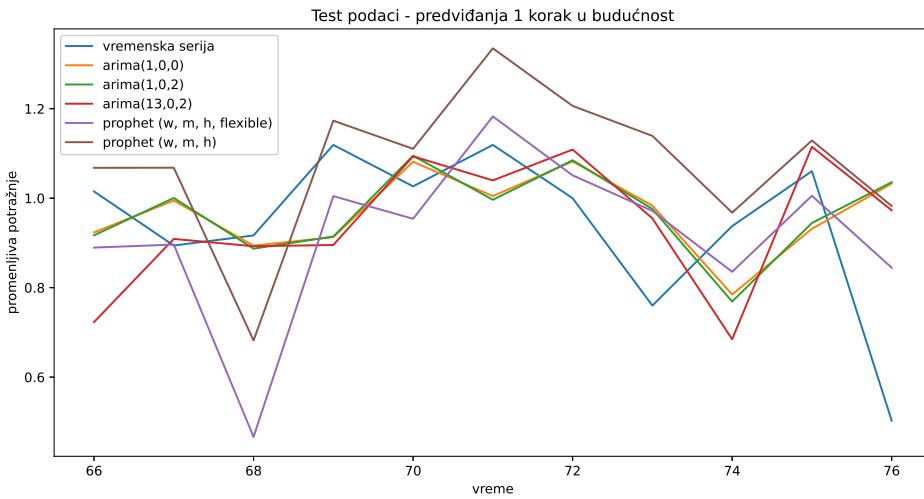
Tabela 3.6: Evaluacione metrike nad test podacima nedeljne vremenske serije.

| model | mae | mse | rmse | smape |
|-----------------------------|----------|----------|----------|----------|
| prophet (w, m, h) | 0.179950 | 0.051546 | 0.227038 | 0.163640 |
| prophet (w, m, h, flexible) | 0.144589 | 0.038063 | 0.195099 | 0.144301 |
| ARIMA(1, 0, 0) | 0.155261 | 0.041368 | 0.203392 | 0.150572 |
| ARIMA(1, 1, 1) | 0.162057 | 0.046475 | 0.215580 | 0.154515 |
| ARIMA(1, 0, 2) | 0.158963 | 0.042111 | 0.205210 | 0.153922 |
| ARIMA(12, 1, 0) | 0.163182 | 0.046216 | 0.214980 | 0.156840 |
| ARIMA(13, 0, 2) | 0.162020 | 0.044062 | 0.209909 | 0.158257 |

kašnjenja (laga) 13, a ostali su isprobani iz radoznalosti. Rezultati AIC metrika za ARIMA modele su prikazani u tabeli 3.5. Iz priložene tabele (3.5) se može primetiti da postoji bolji model po AIC vrednosti od modela koji je funkcija iz *statsmodels* paketa predložila. Na test podacima su evaluirane metrike i ta statistika je prikazana u tabeli 3.6. Pored ARIMA metoda, isprobana su i dva Prophet modela, koja su se istraživanjem pokazala kao najbolja. Na grafiku 3.14 su predstavljene predikcije nad test podacima nekoliko modela.

XGBoost

Kao i kod dnevnih podataka, slično je urađeno i nad nedeljnim podacima. Atributi za XGBoost nad nedeljnim podacima su vrednosti promenljive potražnje na nedeljnog nivou iz prethodnih nekoliko nedelja (konkretno ovde 9). Dodata su i 3 atributa koja predstavljaju praznike u Švedskoj, na isti način kao i za dnevne podatke, samo agregirani na nedeljnog nivou. Metrike evaluacije za ovako dobijeni model su predstavljene u tabeli 3.7, a važnost atributa je predstavljena na grafiku 3.15.



Slika 3.14: Predikcije nad test skupom za nekoliko modela nad nedeljnom vremenom serijom.

Tabela 3.7: Evaluacione metrike nad test podacima nedeljnog XGBoost modela.

| mae | mse | rmse | smape |
|----------|----------|----------|----------|
| 0.146229 | 0.046059 | 0.214613 | 0.138842 |

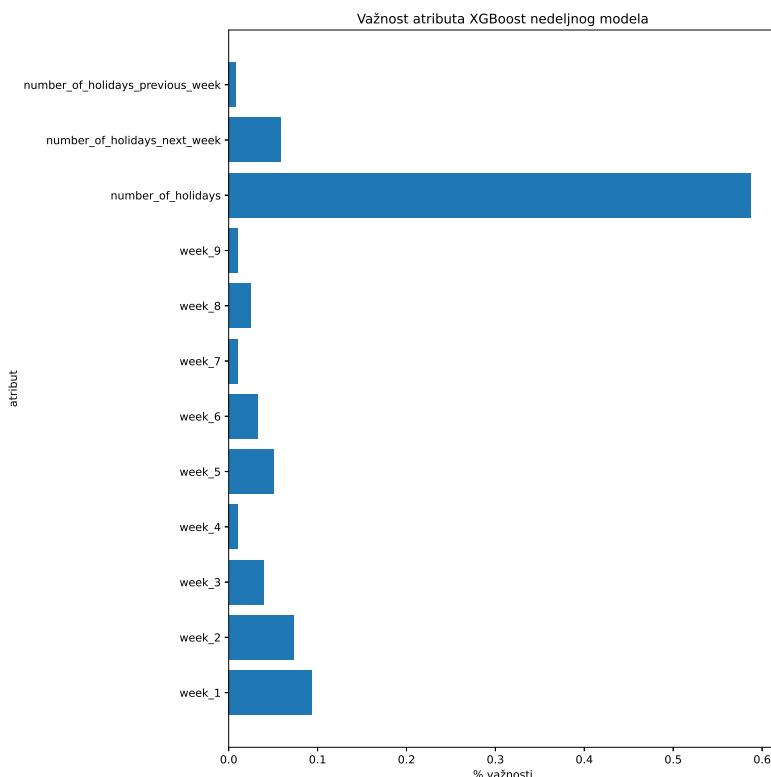
3.4 XGBoost veća granularnost - na nivou automehaničarskih radnji i marki automobila

Kako se dosadašnji tok rada bavio samo podacima na nivou cele države, odnosno agregiranim tako da se predviđa potražnja ciljne promenljive na nivou svih automehaničarskih radnji i marki automobila zajedno, cilj ove sekcije je da se predviđanje dovede do sitnijeg nivoa. ARIMA je isprobana za najveću garažu, za podatke na dnevnom nivou, ali se nije pokazala zadovoljavajuće. Problem je što dosta podataka fali za mnoge garaže ili su vrednosti atributa koji se predviđa dosta niske, tako da vremenska serija možda nije najpogodnije rešenje. U ovoj sekciji konkretno je nivo posmatranog grupisanja po svakoj marki, radnji i danu. Ovim istraživanjem uočeni su neki problemi ovakvog pristupa i oni će biti izloženi u sekciji diskusije, a u nastavku je predstavljen tok analize i rezultati iste.

Isprobane su 4 predikcije nad 4 različita skupa podataka, različita po načinu grupisanja, i to:

- grupisanje marki automobila po različitim garažama na nedeljnom nivou

GLAVA 3. PRIKAZ RADA METODA I REZULTATI



Slika 3.15: Bitnost atributa kod XGBoost nedeljnog modela.

- grupisanje marki automobila po različitim garažama na mesečnom nivou
- grupisanje kreiranih klasa automobila po različitim garažama na nedeljnem nivou
- grupisanje kreiranih klasa automobila po različitim garažama na mesečnom nivou

Deo skupa podataka koji je korišćen za model na nedeljnem nivou po automobilskim markama, prikazan je na slici 1.2. Model sa klasama automobila je uzimao u obzir 4 klase automobila. Te klase su HIGH END, MEDIUM HIGH, MEDIUM LOW, LOW END. Na primer: marka Porsche pripada HIGH END klasi, marka Volvo MEDIUM HIGH klasi, marka Renault MEDIUM LOW, a Kia LOW END itd. Klasifikovanje je urađeno prema predlozima koji su došli iz kompanije. Neki atributi modela su numerički, a neki kategorički i te kategoričke je bilo potrebno enkodirati. Numerički atributi korišćeni su: *x_unit_cost*, *number_of_competitors*, *reachable_population*, a kategorički koji su enkodirani u zavisnosti da li su deo skupa podataka koji se

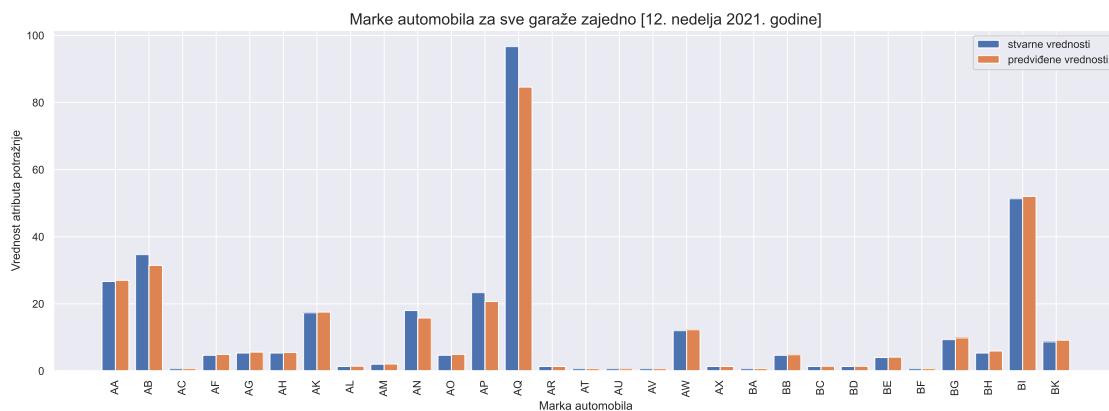
GLAVA 3. PRIKAZ RADA METODA I REZULTATI

Tabela 3.8: XGBoost metrike evaluacije za 4 modela

| model | mae | mse | rmse | smape |
|----------------------------|----------|----------|----------|----------|
| year_week_garage_category | 0.304421 | 0.256627 | 0.506584 | 0.314601 |
| year_month_garage_make | 0.175464 | 0.111608 | 0.334077 | 0.200943 |
| year_week_garage_make | 0.182637 | 0.128826 | 0.358923 | 0.207016 |
| year_month_garage_category | 0.330662 | 0.278295 | 0.527537 | 0.315657 |

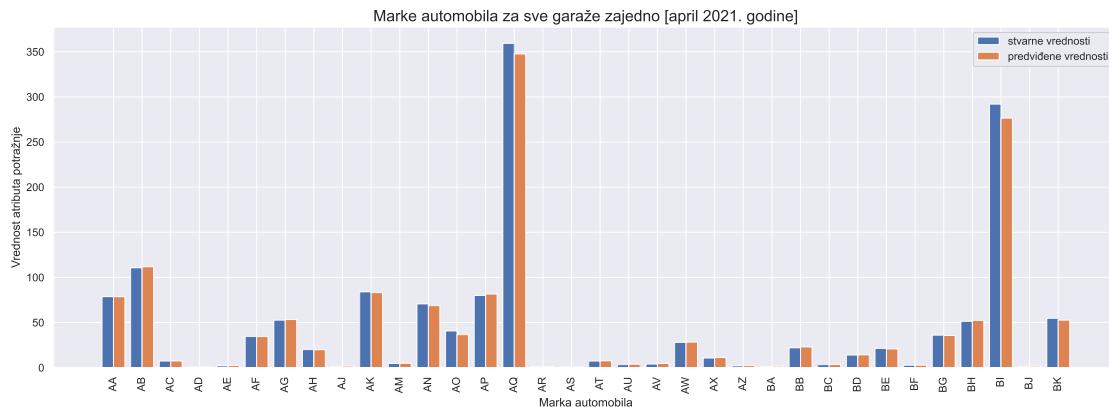
razmatra su: *garage*, *year*, *week_of_year*, *year_week*, *month*, *vehicle_make*, *vehicle_category*. Korišćeno enkodiranje je nadogradnja na Target Encoding. Objasnjeno je u sekciji 2.5, implementirano u okviru rada, a smatra se da može doneti stabilnije rezultate u budućnosti, jer kôd garaža može da bude dosta veliki što ne bi bilo pogodno za neko drugo enkodiranje i unošenje dodatnih kolona.

Metrike evaluacije za dobijene modele su predstavljene tabelom 3.8. U nazivu modela *make* označava da je grupisanje podataka rađeno po marki automobila, a *category* da je grupisanje podataka rađeno po klasi kojoj marka pripada. Na histogramima ispod su prikazane vrednosti automobilskih marki grupisane po svim radnjama. Slika 3.16 prikazuje raspodelu nad test setom koji predstavlja predikciju za 12. nedelju 2021. godine, a slika 3.17 prikazuje raspodelu nad test setom koji predstavlja predikciju za april mesec 2021. godine. Testiranje je rađeno samo nad ovim skupovima podataka.



Slika 3.16: Raspodela marki automobila za 12. nedelju 2021. godine.

GLAVA 3. PRIKAZ RADA METODA I REZULTATI



Slika 3.17: Raspodela marki automobila za april mesec 2021. godine.

3.5 Diskusija rezultata

Prethodne sekcije doprinele su da se bolje razumeju trenutno dostupni podaci. Otkriveno je da veliki broj automehaničarskih radnji spada u „manje” radnje, odnosno da mnoge od njih imaju veoma nisku vrednost ciljne promenljive *demand_value* koja je predviđana ili je uopšte nemaju za dosta vremenskih trenutaka. To znači da je za problem vremenskih serija agregacija podataka na neki način bila neophodna i da posmatranje vremenskih serija za svaku automehaničarsku radnju po automobilskim markama još uvek nema značajan broj podataka.

Takođe, potrebno je da se novi izvor podataka (booking portal) sa kog su podaci i prikupljeni, još više aktivira kako bi reprezentacija podataka bolje predstavljala radnje ili da se podaci obogate i svim drugim izvorima podataka koji postoje za sve automehaničarske radnje (što je problem na kom se radi).

Dobra strana istraživanja vremenskih serija i XGBoost metoda na način na koji je posmatran za dnevne i nedeljne podatke, je da se vrednosti prethodnih dana/nedelja mogu ubaciti kao novi veštački atribut u skup atributa za predviđanje metodama mašinskog učenja, jer se pokazalo da postoji neki smisao ubacivanja vrednosti lagova (vrednosti iz prošlosti) i da to može da doprinese modelu da uči. Broj vrednosti lagova bi mogao biti određen Arimom i njenim koeficijentima.

Prophet je pre svega ispitivan zbog ideje da ima podršku da uračuna efekte praznika na uticaj modela. Kroz istraživanje se pokazalo da je bolje ubaciti efekte praznika i da oni mogu da imaju značajne efekte na model. Pre svega je grafik vremenske serije na nedeljnem nivou otkrio značajne padove oko nekih datuma koji potencijalno nastaju zbog praznika.

GLAVA 3. PRIKAZ RADA METODA I REZULTATI

Predstavljeno klasterovanje automobilskih marki je davalо lošije rezultate, tako da je odbačena ideja klasterovanja automobila na predložen način. Potrebno je isprobati klasterovanje po nekom drugom kriterijumu, a možda i klasterovanje radnji treba uzeti u razmatranje, kako bi podaci ipak bili agregirani, ali ne suviše zbog manjka podataka.

Uočen je i problem posmatranja problema kao sasvim regresionog problema kod XGBoost metode. Dolazak određene automobilske marke u neku automehaničarsku radionicu biva zabeležen i model se recimo obučava na takvим podacima, ali generalno regresioni problem neće da kaže da li je određena automobilska marka uopšte došla ili nije u automehaničarsku radnju. Dakle, potencijalno bi prvo trebalo uraditi binarnu klasifikaciju da je model došao ili nije došao i onda predviđati potražnju pomenute ciljne promenljive.

Glava 4

Zaključak i pravci daljeg rada

Kao što je pomenuto u uvodnom delu, problem predviđanja potražnje je široko zastupljen problem. U automobilskoj industriji precizno predviđanje može da dovede do optimizacije mnogih poslova, a samim tim i do uštede resursa, odnosno povećanja profita. Poslovi koji se mogu optimizovati u ovom konkretnom slučaju su naručivanja potrebne i dovoljne količine automobilskih delova, optimalnog distribuiranja tih delova do automehaničarskih radnji, pametno postavljanje popusta na popravke određenih automobilskih marki kako bi se povećali prihodi, a iskoristili neki zaostali delovi itd.

Ovaj rad je pokušao da načne novi problem koji je predstavljen od strane industrijskog sveta. Rad nad realnim podacima je kao i uvek dosta izazovan, posebno u slučaju kada podataka ima jako malo i u ovom slučaju gde je situacija da je izvor za podatke krenuo sa radom kao inovativna i nova ideja tek krajem 2019. godine. Nakon samog pokretanja desila se velika pandemija virusa i to je svakako uticalo da svi ovi dostupni podaci budu na neki način neobični. To će biti interesantno istraživati u godinama koje tek dolaze, sa povećanjem količine podataka. Takođe, kada se prikupe i podaci iz drugih izvora koji su vezani za sve automehaničarske radnje, mogao bi da se ispita uticaj booking portala na globalnu sliku rada kompanije. Istraživanje je pokazalo da modeli mogu dobro da se ponašaju na agregiranom nivou (recimo nivou cele države) ili na nivou predviđanja velikih automehaničarskih radnji za koje postoje dovoljne količine podataka. Za male radnje postoji veliki problem sa nepostojećim informacijama vezanim za marke automobila, tako da bi možda neki drugi problem (klasifikacija+regresija) mogao da pomogne. Takođe, bitno je napomenuti da ono što je prema podacima sa booking portala mala ili velika radnja, ne mora da bude realna slika. To može da bude samo usled dostupnih podataka ili usled toga

GLAVA 4. ZAKLJUČAK I PRAVCI DALJEG RADA

što se neka velika radnja tek skoro priključila ozbilnjijem korišćenju ovog načina rezervisanja popravki, pa deluje kao da je mala po količini ciljne promenljive. To se može nadomestiti dobijanjem pristupa drugim izvorima podataka u ovoj kompaniji.

Istraživanje je otvorilo i neke nove ideje. Pre svega sledeći korak bi mogao biti obogaćivanje trenutnih podataka podacima koji su javno dostupni, a vezani za vremensku prognozu. Moguće je za svaki vremenski trenutak i na osnovu lokacija automehaničarskih radnji dodati atribut koji bi predstavljaо svojstva vremenske prognoze, kako bi se uočilo da li zima/leto imaju uticaj na to kada ljudi popravljaju svoje automobile. Takođe, više informacija o samim korisnicima/garažama/automobilima bi doprinelo boljem klasterovanju podataka. Informacije o svakom automobilu koji je bio na popravci sadrže godinu proizvodnje, tako da se može gledati u pravcu da se obogaćivanje podataka vrši proširivanjem aproksimirane cene tog automobila i nekakve grupacije na osnovu tih informacija. Korišćenje Google Mapa umesto Open Street Mapa može doprineti preciznijem stanju, a možda i dobijanje javnih statističkih informacija o konkurentnim radnjama može da se uzme u obzir. Uzimanje više ličnih informacija korisnika ili više informacija o automobilima (pređena kilometraža, broj ukupnih vlasnika, broj oštećenja itd.) prilikom unosa rezervacije na booking portal, bi takođe moglo biti iskorišćeno u svrhe boljeg klasterovanja, a samim tim i predviđanja.

Bibliografija

- [1] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- [2] Boram Choi and Jong Hwan Suh. Forecasting spare parts demand of military aircraft: Comparisons of data mining techniques and managerial features from the case of south korea. *Sustainability*, 12(15):6045, 2020.
- [3] European Commission. European Commission Climate action, road-map for moving to a low-carbon economy in 2050, 2013. on-line at: https://ec.europa.eu/clima/sites/clima/files/strategies/2050/docs/roadmap_fact_sheet_en.pdf.
- [4] M Faccio, F Sgarbossa, and A Callegaro. Forecasting methods for spare parts demand. *Italy: Universita'Degli Studi di Padova*, page 6, 2010.
- [5] Robby Henkelmann. A deep learning based approach for automotive spare part demand forecasting. *Otto von Guericke Universität Magdeburg*, 2018.
- [6] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.
- [7] Irappa Madabhavi, Malay Sarkar, and Nagaveni Kadakol. Covid-19: a review. *Monaldi Archives for Chest Disease*, 90(2), 2020.
- [8] John Mello et al. Demand and supply integration: The key to world-class demand forecasting by mark a. moon. *Foresight: The International Journal of Applied Forecasting*, (31):35–37, 2013.
- [9] Andelka Zečević Mladen Nikolić. Mašinsko učenje - skripta, 2019. on-line at: <http://ml.matf.bg.ac.rs/readings/ml.pdf>.

BIBLIOGRAFIJA

- [10] Florian Pargent, Bernd Bischl, and Janek Thomas. *A benchmark experiment on how to encode categorical features in predictive modeling*. PhD thesis, M. Sc. Thesis, Ludwig-Maximilians–Universitat Munchen, pp12, 2019.
- [11] AM Saravanan, SP Anbuudayasanakar, and P Arul William David. Forecasting techniques for sales of spare parts. *International Journal of Recent Technology and Engineering*, 8(3):27–30, 2019.
- [12] Facebook Open Source. Prophet, 2021. on-line at: <https://facebook.github.io/prophet/>.
- [13] Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.
- [14] LF Tratar and E Strmčnik. Forecasting methods in engineering. In *IOP Conference Series: Materials Science and Engineering*, volume 657, page 012027. IOP Publishing, 2019.
- [15] CA González Vargas and M Elizondo Cortés. Automobile spare-parts forecasting: A comparative study of time series methods. *International Journal of Automotive and Mechanical Engineering*, 14:3898–3912, 2017.
- [16] Jafri Zulkepli, Chan Hwa Fong, and Norhaslinda Zainal Abidin. Demand forecasting for automotive sector in malaysia by system dynamics approach. In *AIP Conference Proceedings*, volume 1691, page 030031. AIP Publishing LLC, 2015.
- [17] Emir Zunic, Kemal Korjenic, Kerim Hodzic, and Dzenana Donko. Application of facebook’s prophet algorithm for successful sales forecasting based on real-world data. *arXiv preprint arXiv:2005.07575*, 2020.

Biografija autora

Andelka Milovanović je rođena u Valjevu na datum koji predstavlja aproksimativnu vrednost broja π , dakle 22. jula 1996. godine. Išla je u Osnovnu školu „Sestre Ilić“ u Valjevu i u Muzičku školu „Živorad Grbić“, odsek za flautu. Nakon završene osnovne škole, kao najbolji čak svoje generacije, upisala je specijalizovano-matematičko odjeljenje u Valjevskoj gimnaziji. Paralelno sa gimnazijom, pohađala je srednju muzičku školu za flautu nekoliko meseci, dok se nije javilo interesovanje za neke druge oblasti.

Tokom srednje škole bila je polaznica seminara astronomije u Istraživačkoj stanicici Petnica sve četiri godine i vodila je astronomsku grupu u Društву istraživača „Vladimir Mandić-Manda“ u Valjevu. Organizovala je aktivnosti poput akcija za posmatranje meteorskih rojeva i drugih nebeskih objekata, pomračenja Sunca i Meseča, Sata za našu planetu, zimskih i letnjih astronomskih škola za učenike osnovnih i srednjih škola, raznih naučno-popularnih predavanja i mnoge druge. Fokus tokom srednje škole joj je uglavnom bio na vaannastavnim aktivnostima, tako da je bila jedan od organizatora prvog Festivala nauke u gimnaziji, volontirala je na događajima Centra za promociju nauke i učestvovala na Astro-vikendu u Domu Omladine Beograda.

Najinteresantnija oblast u srednjoj školi joj je bilo programiranje, tako da odlučuje da upiše smer Informatike na Matematičkom fakultetu u Beogradu. Nakon 2. godine fakulteta dobija punu stipendiju Američke Vlade za jednosemestralnu studentsku razmenu „Global UGRAD“, na Univerzitetu na Floridi - Florida Gulf Coast University. Nakon godinu dana pauze na Matematičkom fakultetu, vraća se svojim studijama. Na početku 4. godine studija dobija besplatnu kartu od fakulteta za učešće na PyCon konferenciji u Beogradu, koja je dovodi do osvajanja 1. mesta na lokalnom i globalnom hakatonu (TADHack 2019) zajedno sa svoja dva drugara. Iste godine je provela 1. semestar u ulozi studenta demonstratora na kursu Računarske grafike, a 2. semestar u ulozi praktikanta u Microsoft Development Centru Srbije.

Nakon uspešno završenih osnovnih studija na Matematičkom fakultetu, Andelka nastavlja obrazovanje i upisuje master studije na istom smeru, pokušavajući da otkrije koja oblast je najviše interesuje. Iste godine dobija diplomu Društva istraživača, kao potvrdu velikog zalaganja za popularizaciju nauke i samog društva tokom srednjoškolskog perioda.