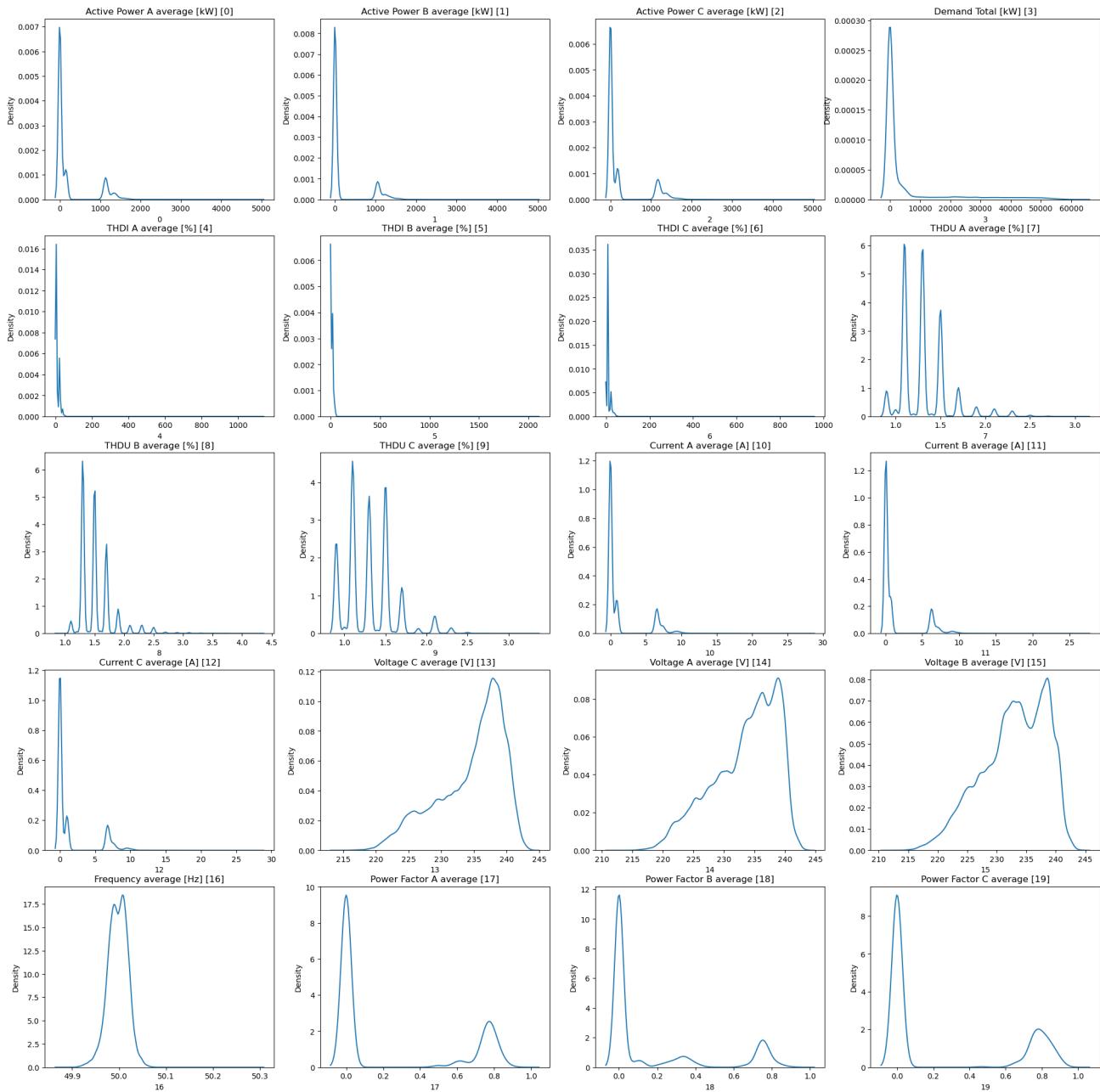
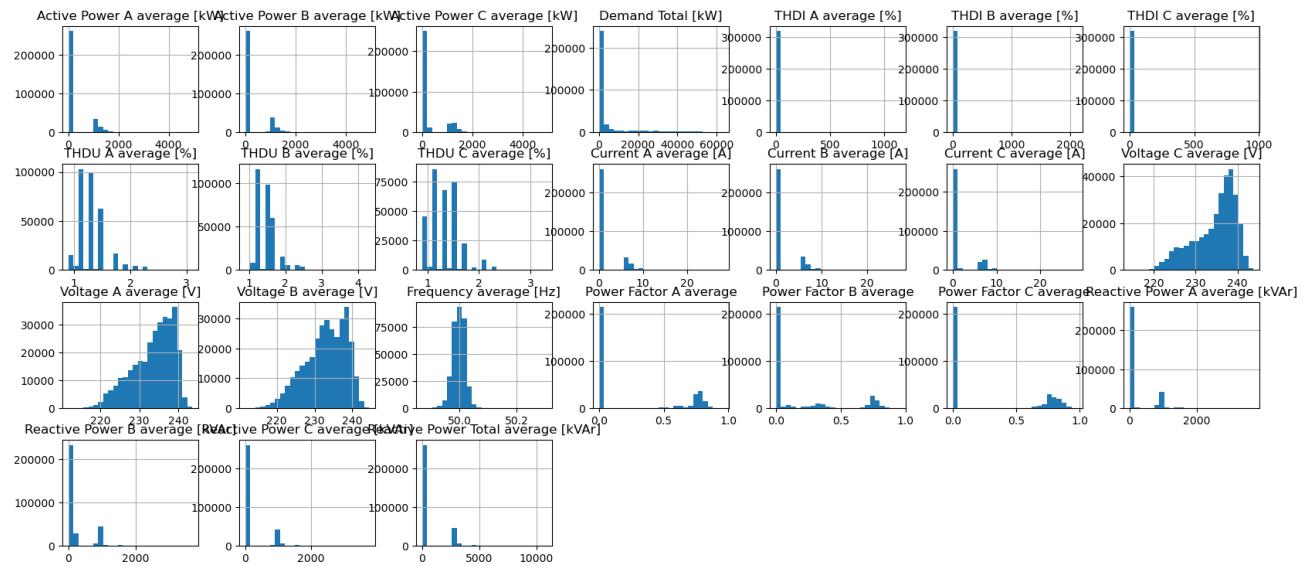


1. Istrazivačka analiza podataka:

Dataset sadrži podatke o energetskim parametrima. Sastoji se od 25 kolona, tj. atributa (uključujući i timestamp atribut) i 318556 vrsta, tj. opservacija. Svaka vrsta predstavlja skup vrednosti energetskih parametara u jednom vremenskom trenutku (timestamp atribut). Podaci su snimani u periodu od 4 dana (2022-11-07 do 2022-11-10). Osim timestamp atributa koji je tipa object, svi atributi su numerickog tipa, konkretno 64-bitni realni brojevi.



Slika 1. Raspodela podataka.



Slika 2. Histogrami.

Budući da je cilj ovog projekta detekcija anomalija, posmatraće se samo period u kom mašina radi, tj. posmatraće se period između 8 i 16 časova za svaki dan. Analizom je utvrđeno da mašina ne radi sve vreme tokom tih 8 sati, već da malo pred kraj prestaje sa radom, tako da se i taj period neće razmatrati u daljoj obradi. Dataset će biti podeljen na 4 dataframe-a koja predstavljaju 4 dana i svaki od njih sadrži opservacije čiji je timestamp između 8 i 16 časova za taj dan i kod kojih je Active Power veći od 0 (pošto u tom slučaju, mašina radi).

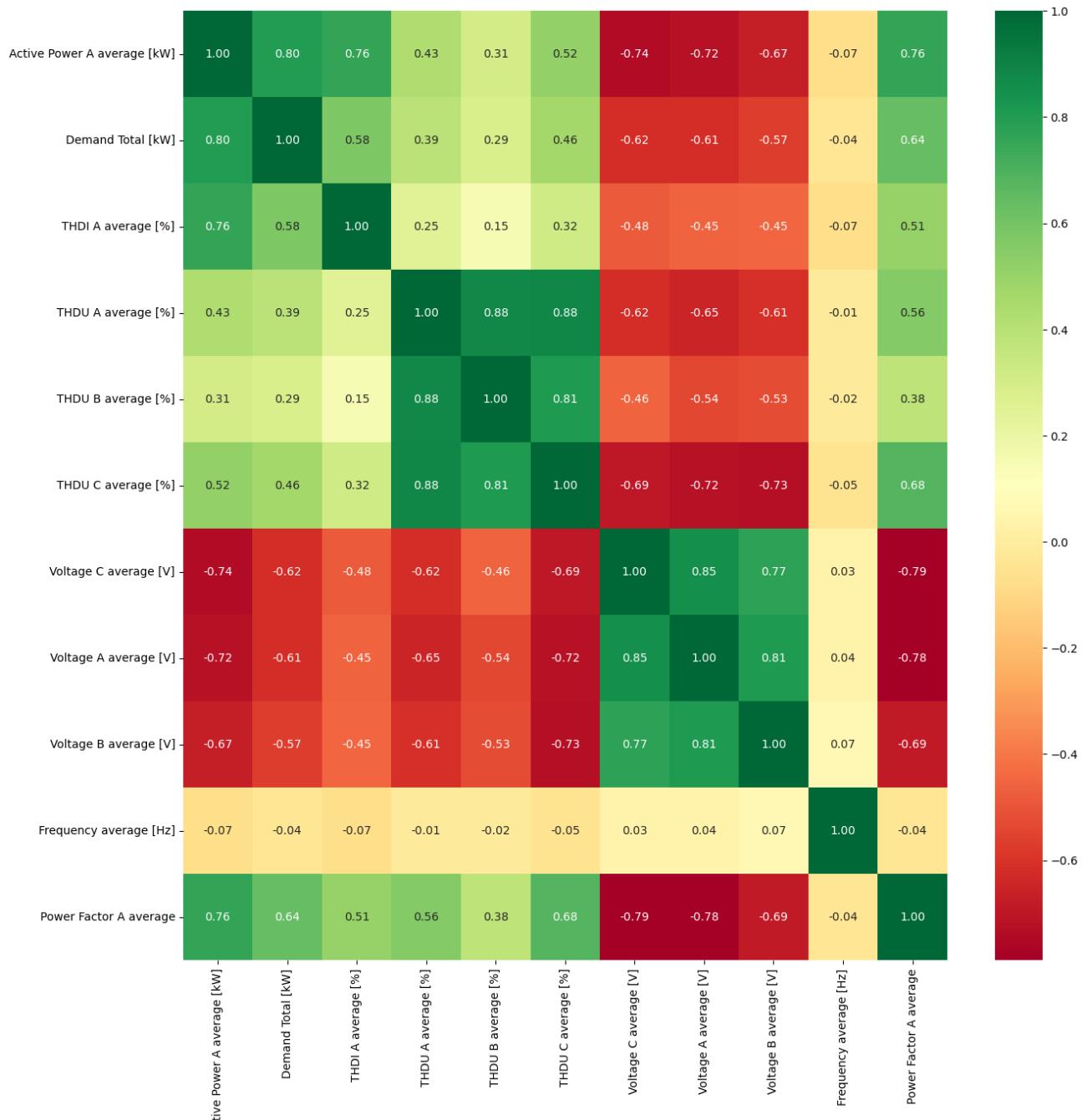
2. Preprocesiranje podataka:

Pri formiranjima dataframe-ova za obradu, bilo je potrebno je obraditi timestamp atribut, tako da on služi za indeksiranje celog početnog dataframe-a. Nakon toga, obavljen je resampling celog dataframe-a tako što se između indeksa prve i poslednje opservacije u dataframe-u, frekvencijom od jedne skunde, ubacuju nedostajuće opservacije metodom ffill (za vrednosti parametara novoubačene vrednosti uzimaju se vrednosti prethodne validne opservacije). Nije bilo duplikata ni null vrednosti. Takođe su uklonjeni THDU A, B i C za koje je utvrđeno da remete rezultate algoritma.

3. Redukcija dimenzionalnosti:

1) Uklanjanje feature-a sa visokom medjusobnom korelacijom:

Uklonjeni su atributi sa međusobnom korelacijom koja je bila veća od 0,95. Uklonjeni atributi su: 'Active Power B average [kW]', 'Active Power C average [kW]', 'THDI B average [%]', 'THDI C average [%]', 'Current A average [A]', 'Current B average [A]', 'Current C average [A]', 'Power Factor B average', 'Power Factor C average', 'Reactive Power A average [kVAr]', 'Reactive Power B average [kVAr]', 'Reactive Power C average [kVAr]', 'Reactive Power Total average [kVAr]'



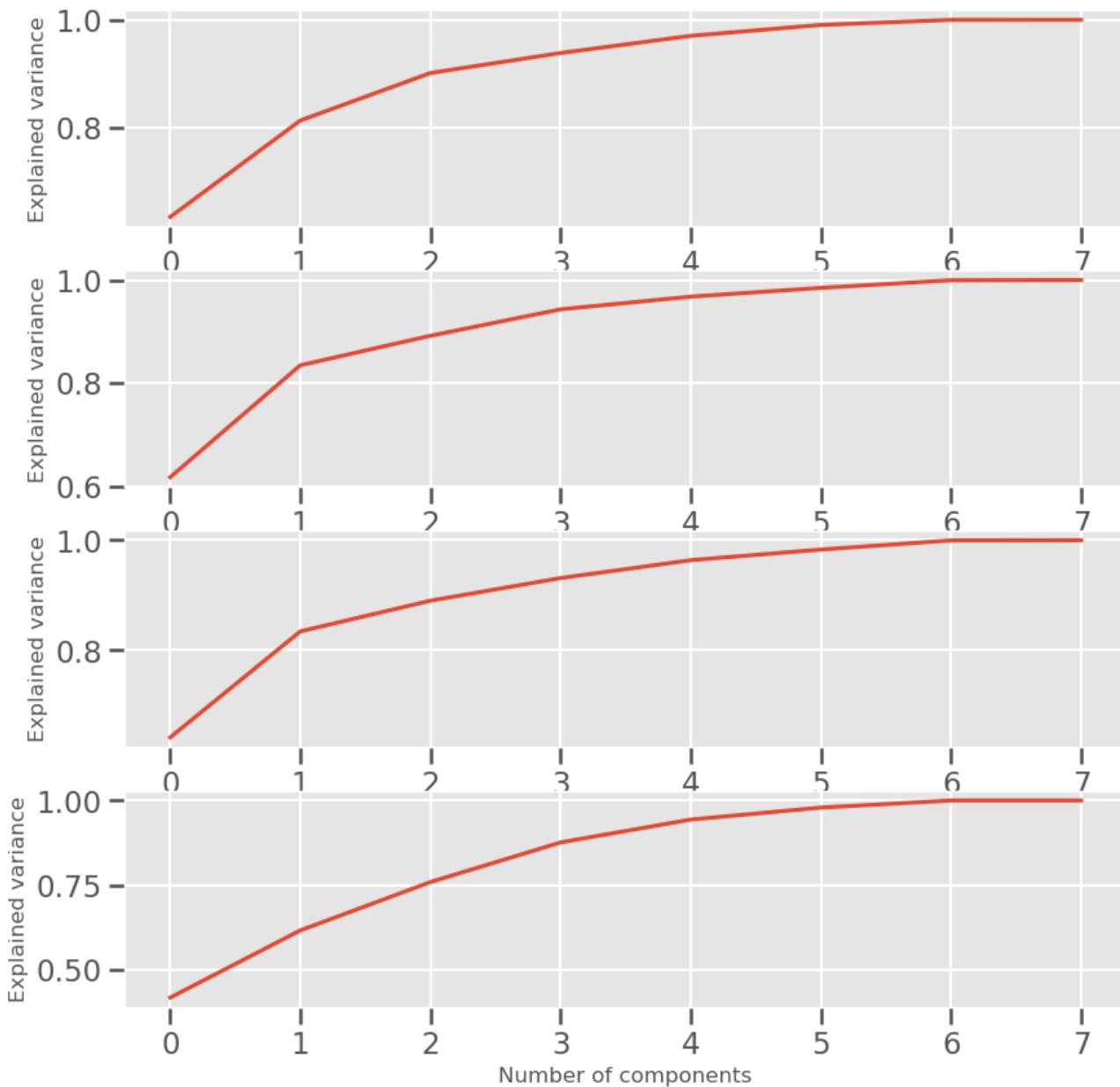
Slika 3. Matrica korelacija.

2) Normalizacija:

Sve vrednosti su normalizovane za potrebe dalje obrade.

3) PCA analiza:

U ovom delu se obavljala podela dataframe-a koja je objašnjena na kraju teze 1. Obavljena je PCA analiza nad svakim od ovih dataframe-ova posebno, i za svaki od njih je formiran novi dataframe koji je sadržao principijalne komponente (po 5 za svaki od 4 dataframe-a).



Slika 4. PCA analiza.

Ovim korakom, formirani su konačni dataframe-ovi koji će se koristiti kao ulaz u algoritme za detekciju anomalija.

4. Detekcija anomalija:

Koristiće se dva algoritma u ovu svrhu: Isolation Forest i Local Outlier Factor. Metodi evaluacije će biti silhouette score i vizuelizacija. Algoritmi će biti pokretani za svaki od četiri dataframe-a (4 dana) posebno i detektovane anomalije će biti analizirane za svaki dan posebno.

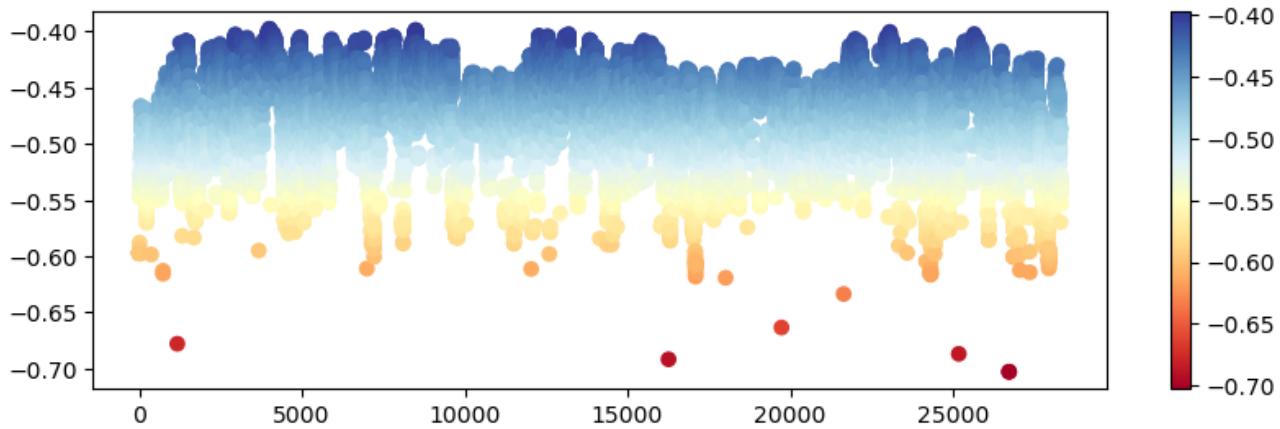
Isolation Forest:

Za svaki dataframe, prikazani su score samples kako bi se stekao uvid u anomaličnost opservacija. Kod IsolationForest algoritma, što je manji score sample za opservaciju, to je ta opservacija više anomalična. Ideja je na osnovu vizuelizacije score samples (uz silhouette score metod evaluaciju) proceniti koji procenat dataframe-a je anomaličan. Uz pomoć ove kontaminacije, može se oprilike videti koje opservacije u dataframe-u su potencijalno anomalične da bi se po mogućstvu primetila neka pravilnost. Za svaki dataframe je primenjen GridSearch algoritam za nalaženje potencijalno optimalnih parametara za IsolationForest algoritam. Parametri od koji je GridSearch birao optimalni su sledeći:

```
'n_estimators':[50, 100, 200, 500, 700]
'max_samples':[50, 100, 200, 400, 700]
'contamination': ['auto', 0.001, 0.01, 0.1]
'bootstrap': [True, False]
'n_jobs': [-1]
'random_state': [42]
```

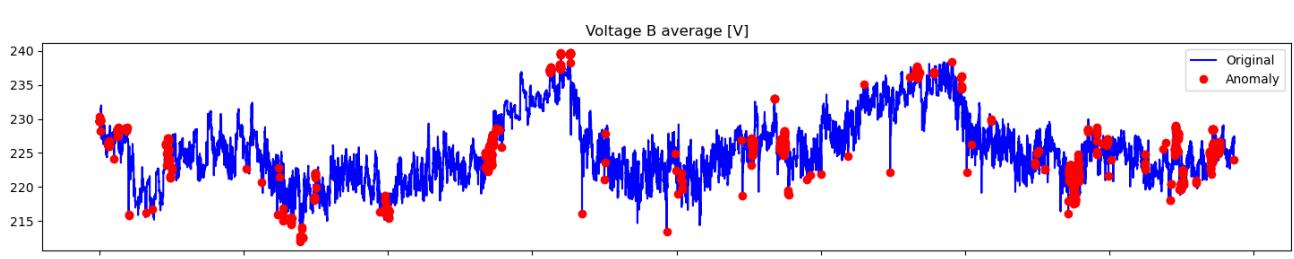
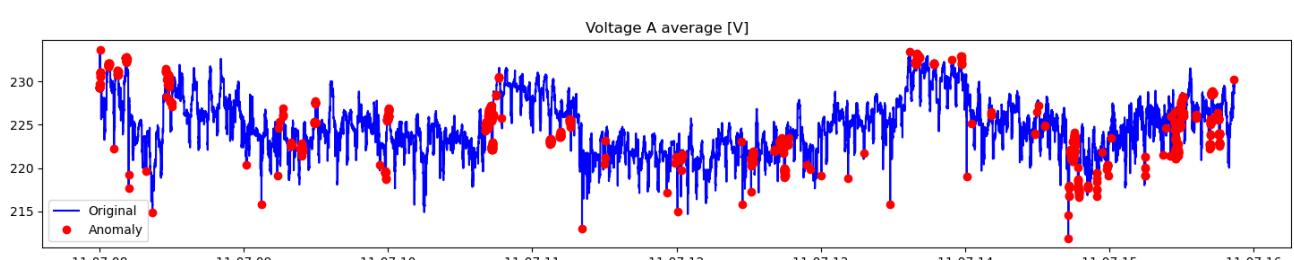
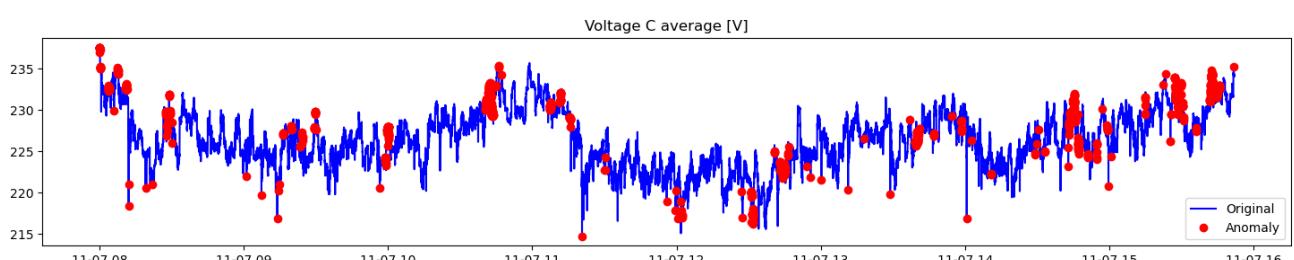
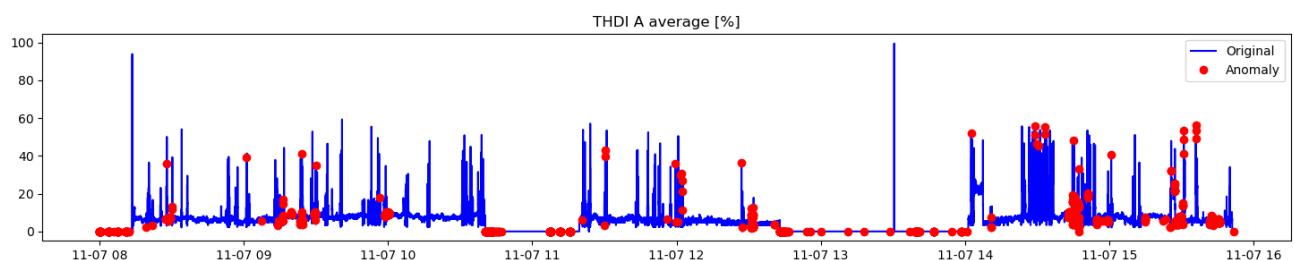
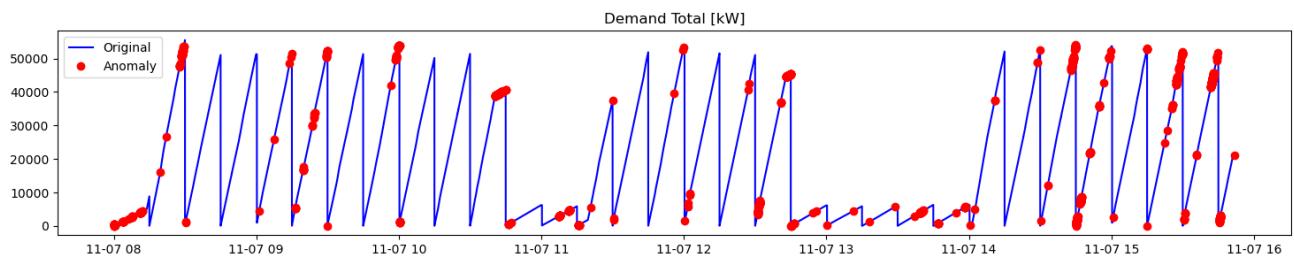
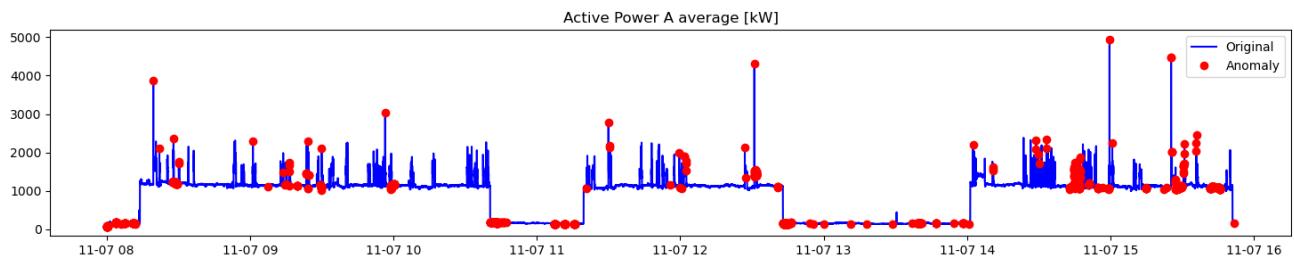
Nakon rezultata GridSearch algoritma, ukoliko bi za neki od parametra koji se ispituju bila izabrana minimalna ili maksimalna vrednost od zadatih, onda bi se taj parameter dodatno ispitivao tako što bi se pokretao algoritam IsolationForest za veću i manju vrednost tog parametra od vrednosti koju je vratio GridSearch. Na kraju se formira skup konačnih parametara IsolationForest algoritma, algoritam se pokreće, analiziraju se dobijeni rezlultati i čuvaju se u csv fajl.

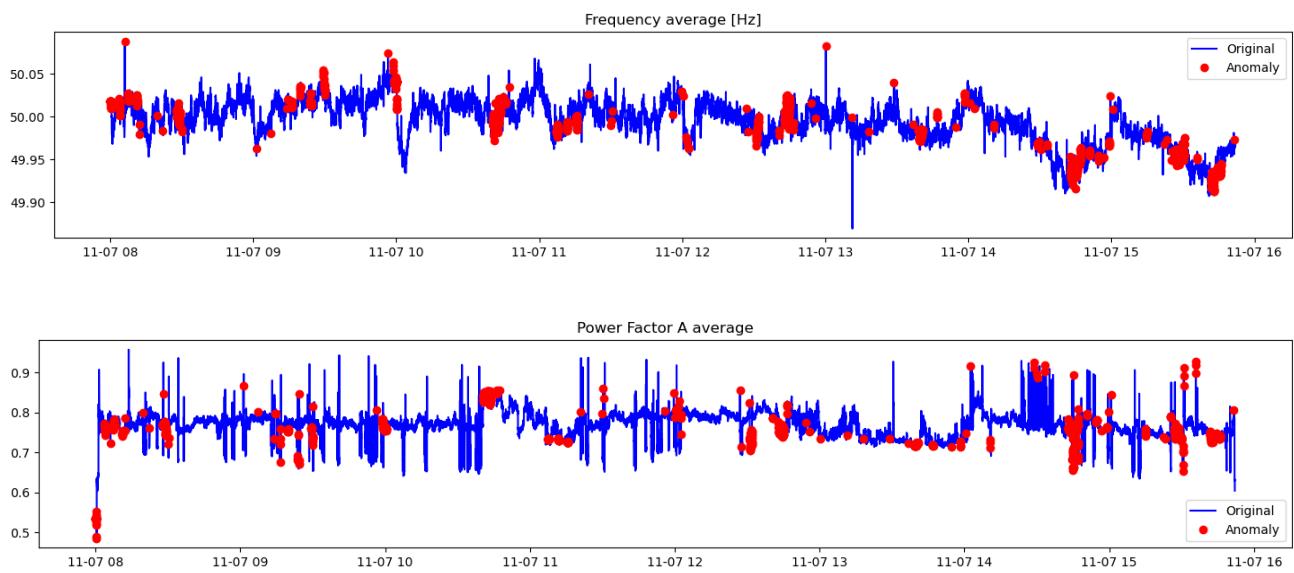
Prvi dan:



Slika 5. Score samples za prvi dan.

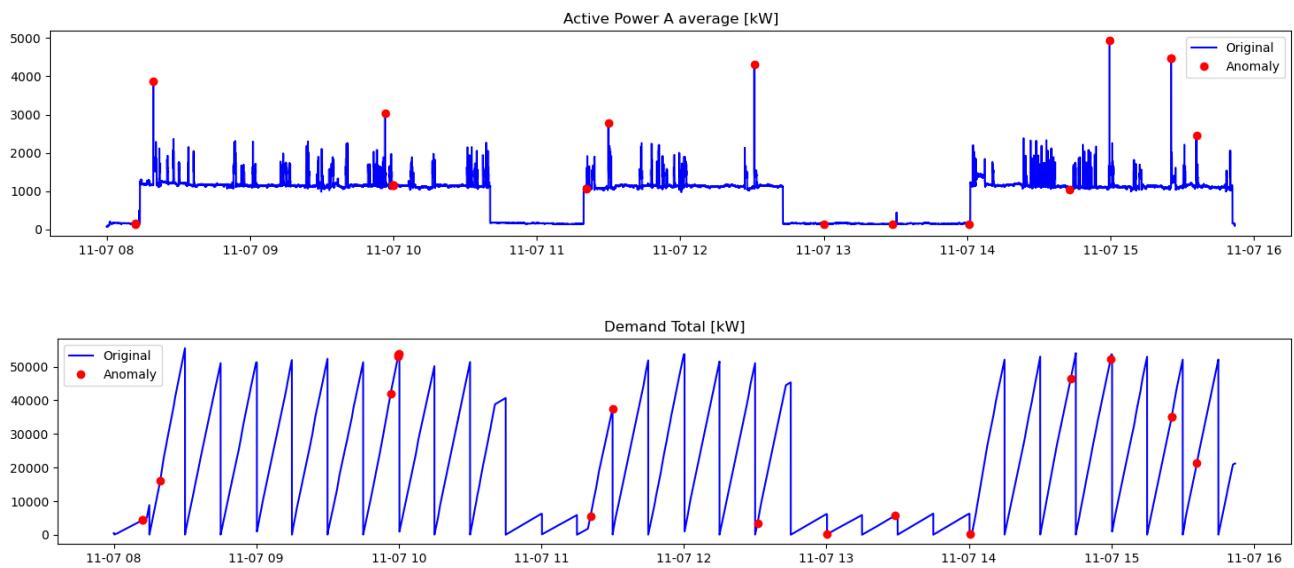
Analizom Score samples, odabrana je početna kontaminacija od 0,0387 (score < -0,55). Rezultati IsolationForest(n_estimators=300, max_samples=40, contamination=0,0387, random_state=42) su sledeći:

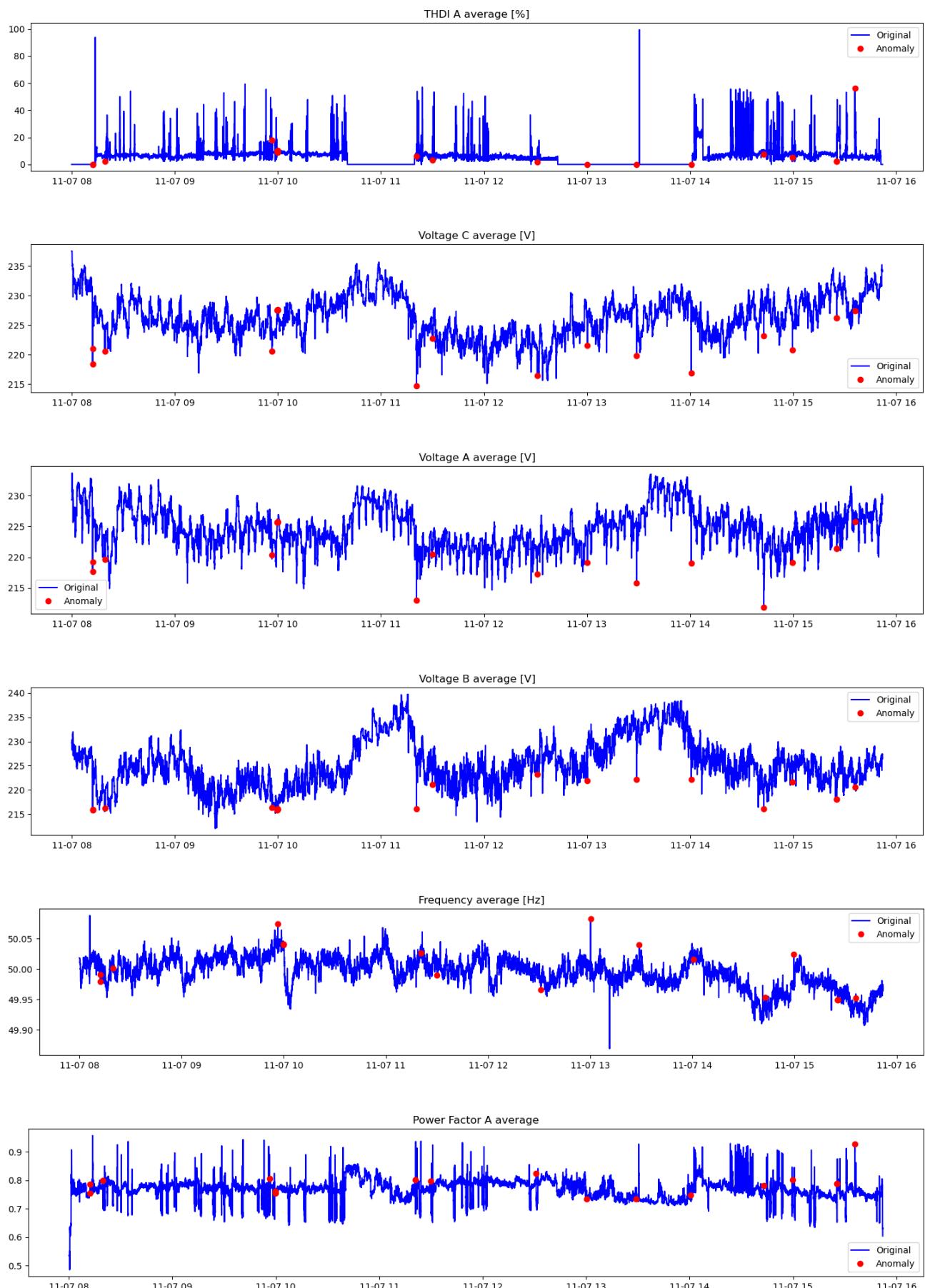




Slika 6. Prvi dan, Isolation Forest ($n_{estimators}=300$, $max_samples=40$, $contamination=0.038764342453662844$). Kontaminacija je prevelika, treba pronaći odgovarajuće parametre IsolationForest algoritma putem GridSerach-a. Međutim, čak i sa toliko visokom kontaminacijom, 2 pika najviših vrednosti u THDI A average nisu detektovani.

GridSearch je vratio sledeće parametre kao najoptimalnije: $n_{estimators}=500$, $max_samples=700$, $contamination= 0.001$, $bootstrap=False$, $random_state=42$. Budući da su vrednosti $max_samples$ i $contamination$ parametra najveća i najmanja vrednost, respektivno, među onima koje su predate GridSearch algoritmu, izvršena je provera silhouette score za $max_samples$ vrednosti 800 i 900, a odmah posle toga i za $contamination$ vrednosti 0.005 i 0.0007. Provera je pokazala da parametri koje je vratio GridSearch i dalje daju najbolje rezultate.



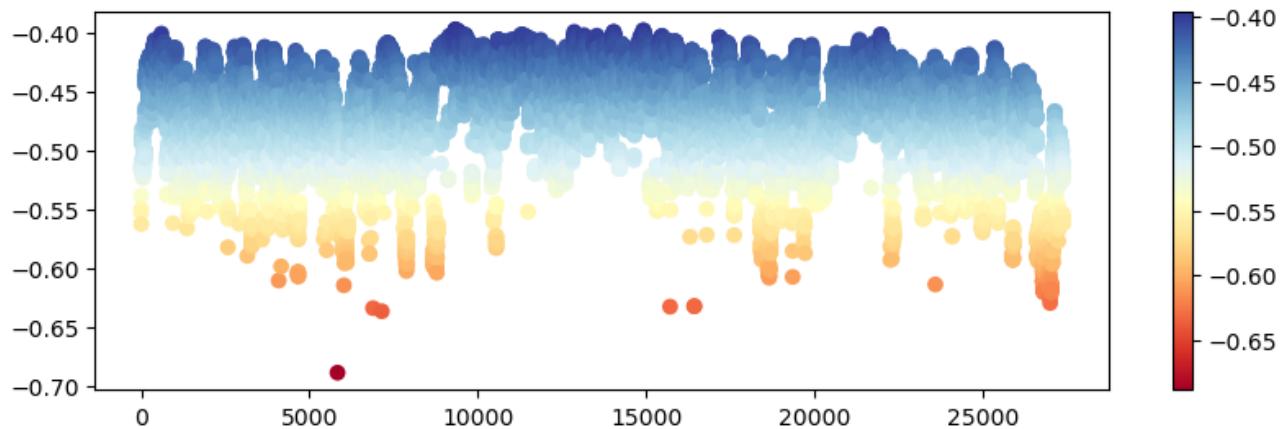


Slika 7. Prvi dan, Isolation Forest ($n_{estimators}=500$, $max_samples=700$, $contamination= 0.001$,

bootstrap=False, random_state=42).

1. Detektovani su pikovi najvisih vrednosti u Active Power A average.
2. Kod Voltage A, B i C su uglavnom detektovane vrednosti niskih vrednosti.

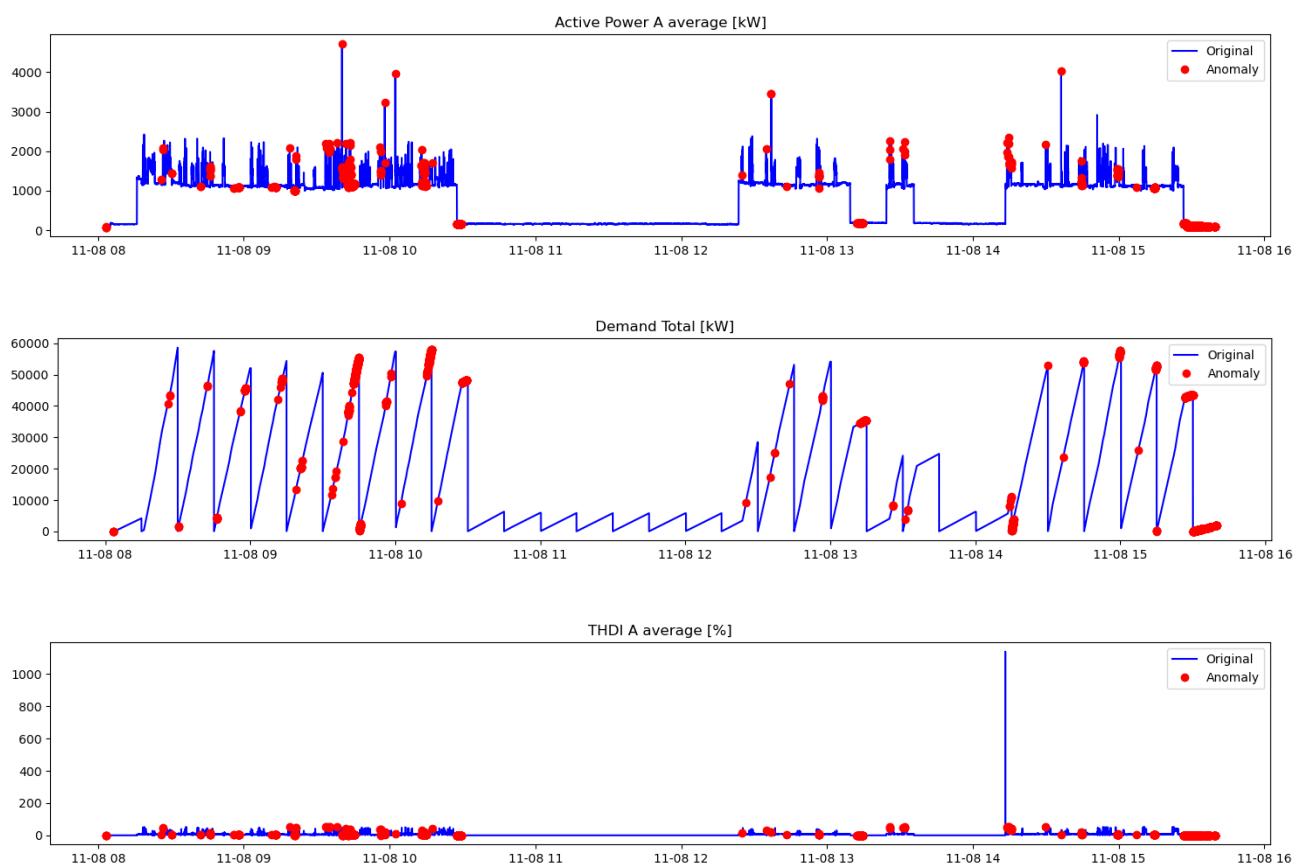
Drugi dan:

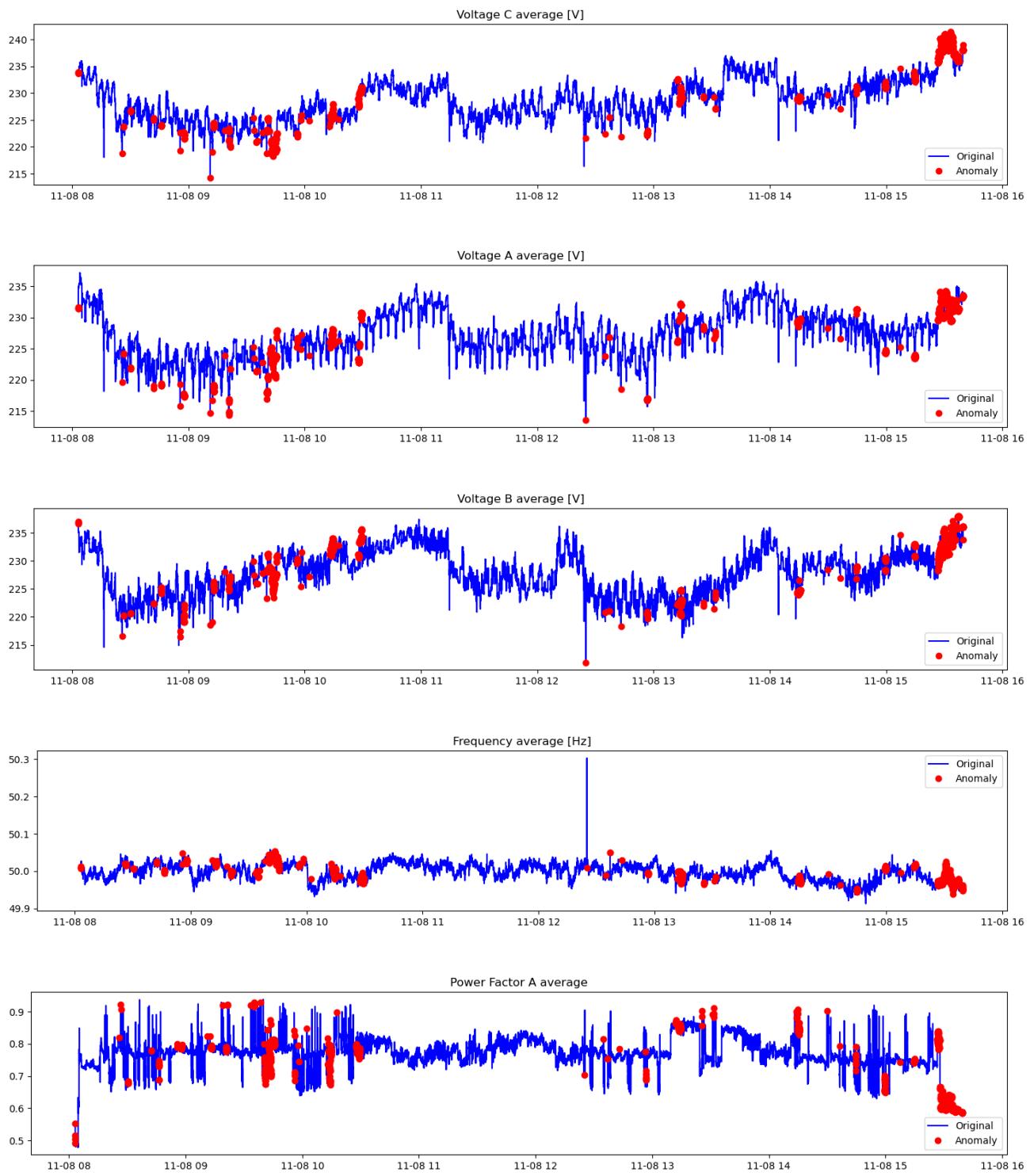


Slika 8. Score samples za drugi dan.

Analizom score samples je odabrana početna kontaminacija od 0.041729 (scores < -0.56).

IsolationForest(n_estimators=300, max_samples=40, contamination=0.041729, random_state=42) je dao silhouette score 0.26271255.



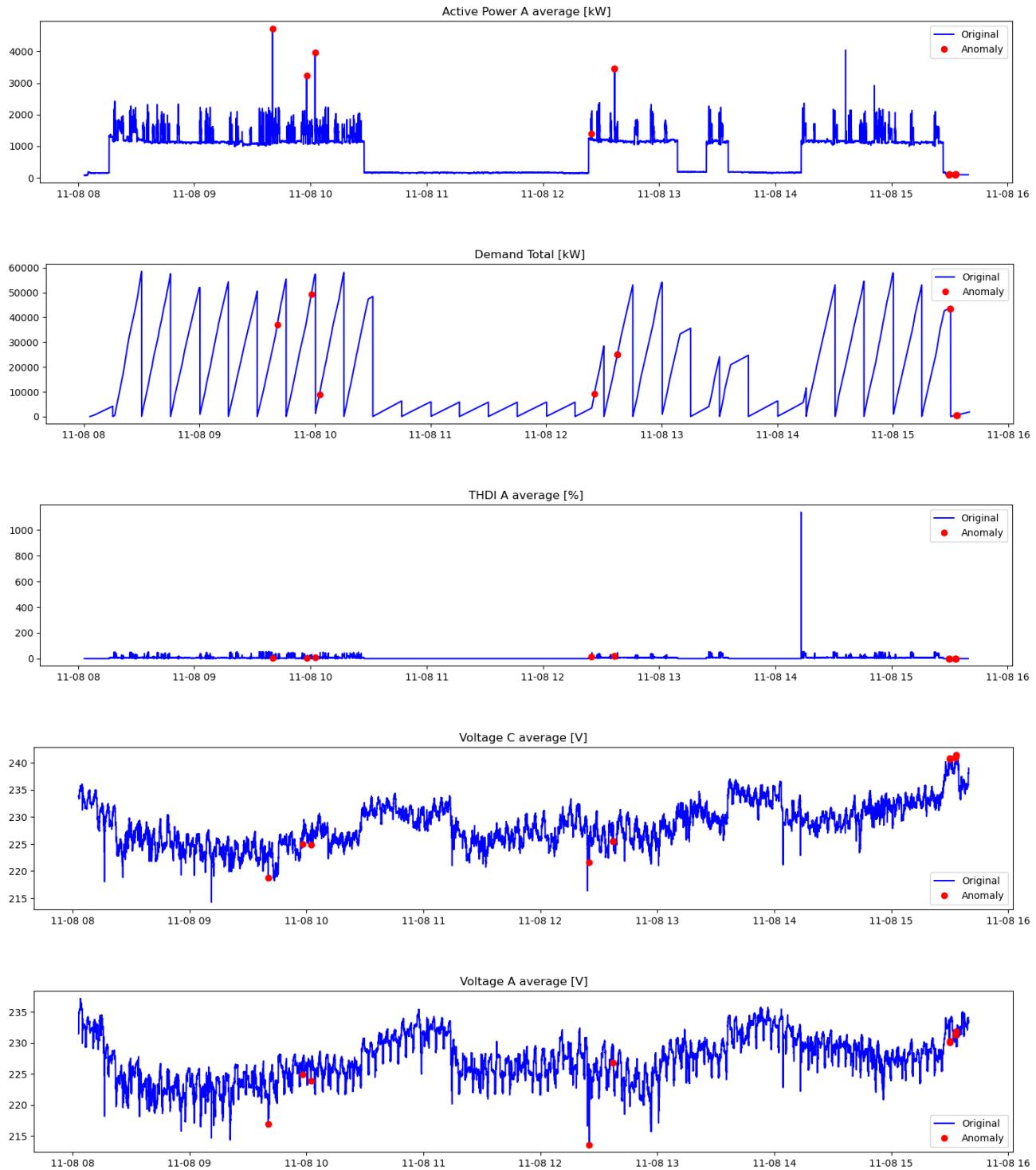


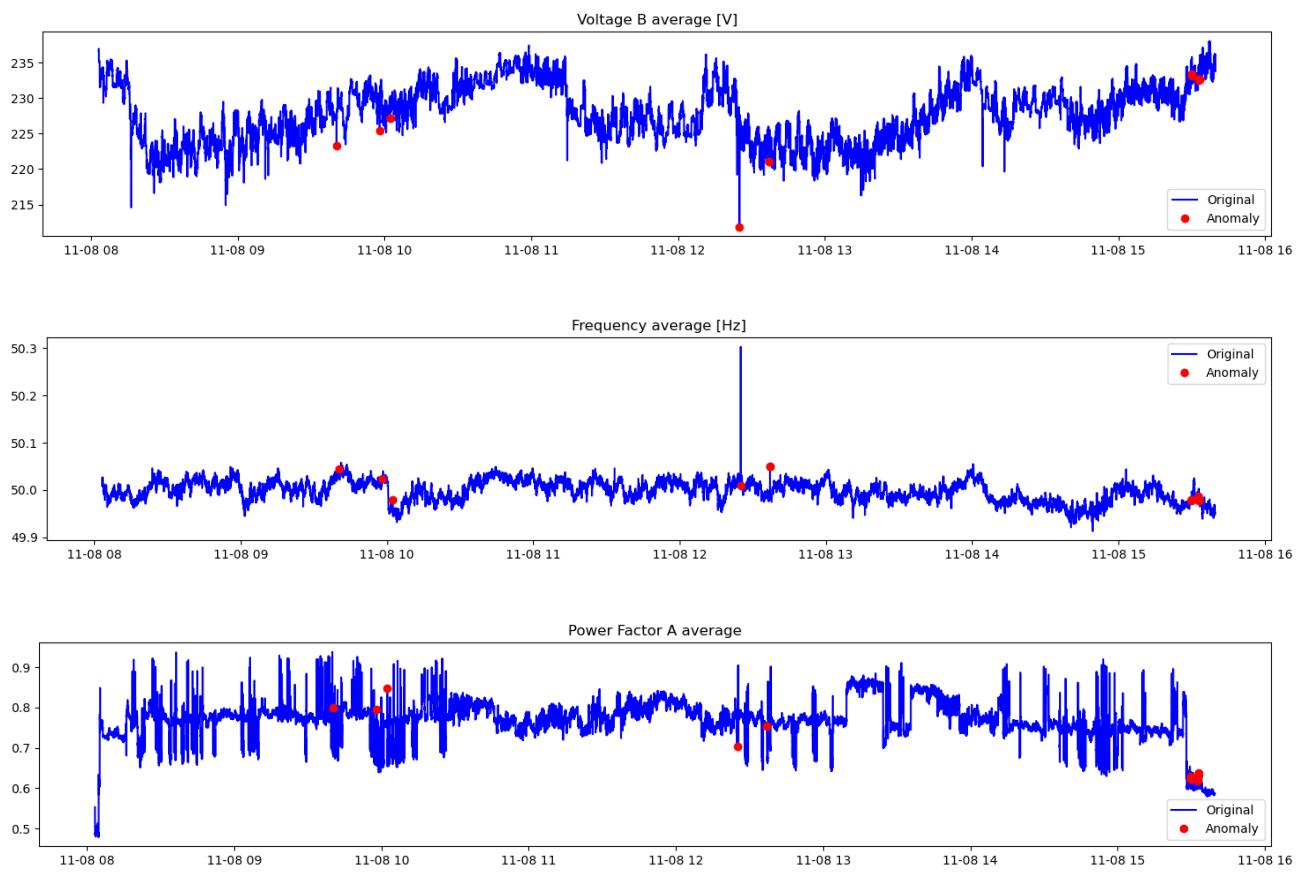
Slika 9. Drugi dan, Isolation Forest (n_estimators=300, max_samples=100, contamination=0.041729).

1. Nisu detektovani pikovi najviših vrednosti u THDI A average i Frequency average, uprkos očigledno previsokoj kontaminaciji.
2. Detektovani pikovi najviših vrednosti u Active Power A average.
3. Gusta koncentracija detektovanih anomalija pred kraj radnog vremena (pred kraj intervala od 15h do 16h).

Sledi pokretanje GridSearch algoritma.

GridSearch je vratio sledeće parametre kao najoptimalnije: n_estimators=50, max_samples=400, contamination= 0.001, bootstrap=True, random_state=42. Budući da je vrednost contamination parametra najmanja vrednost među onima koje su predate GridSearch algoritmu, izvršena je provera silhouette score za contamination vrednosti 0.005 i 0.0007. Provera je pokazala da parametri koje je vratio GridSearch i dalje daju najbolje rezultate.

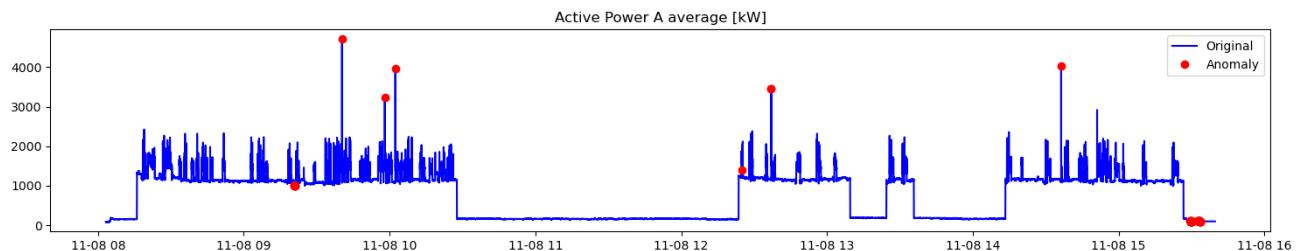


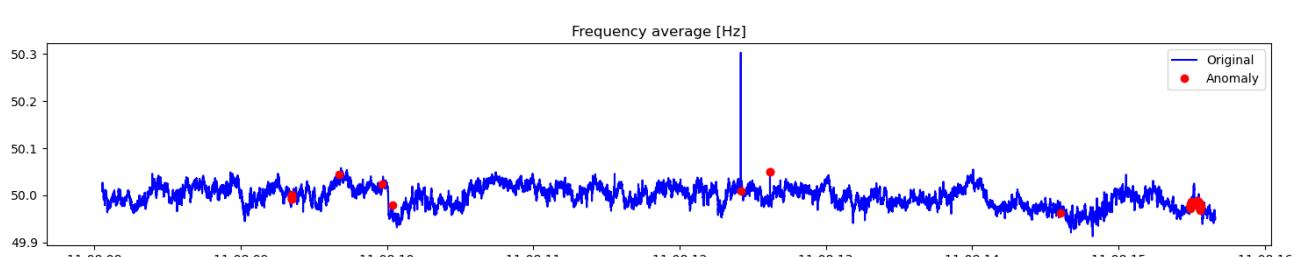
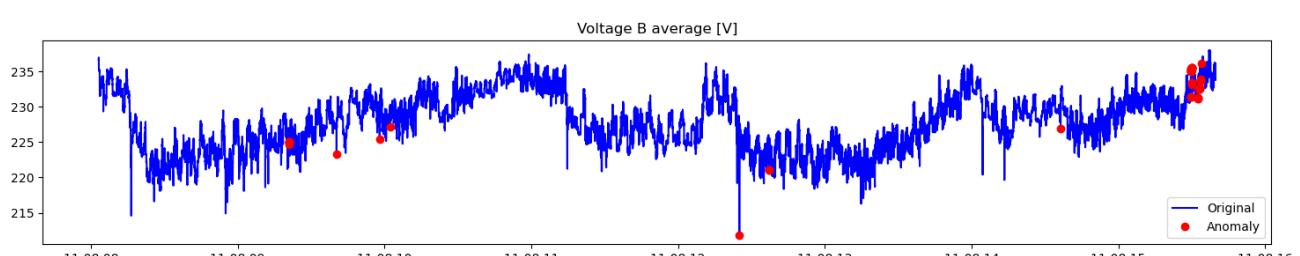
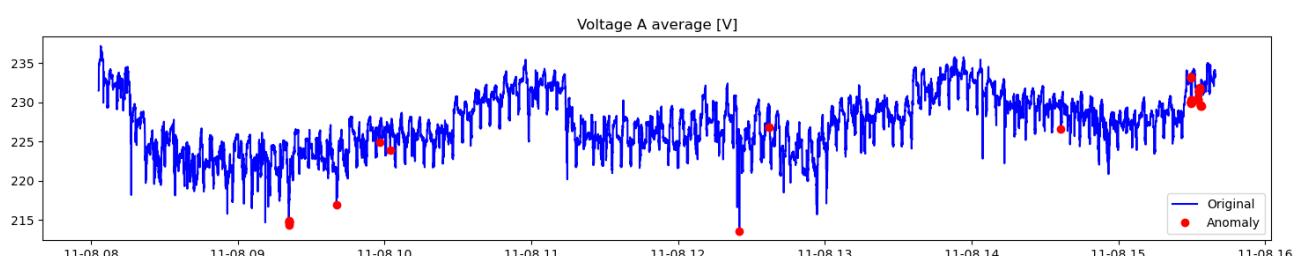
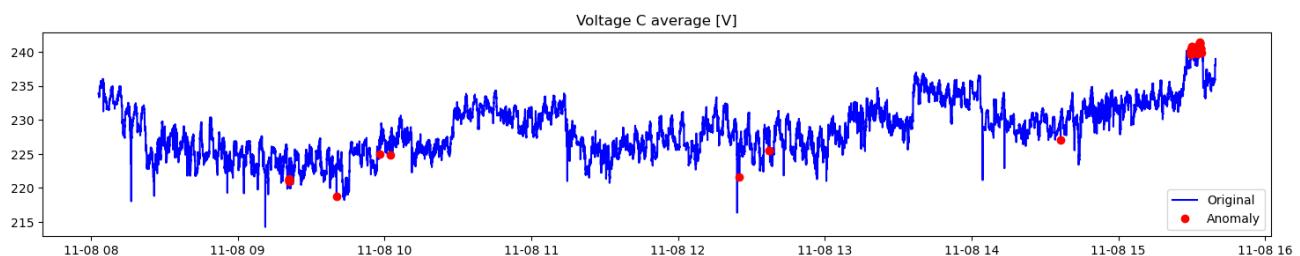
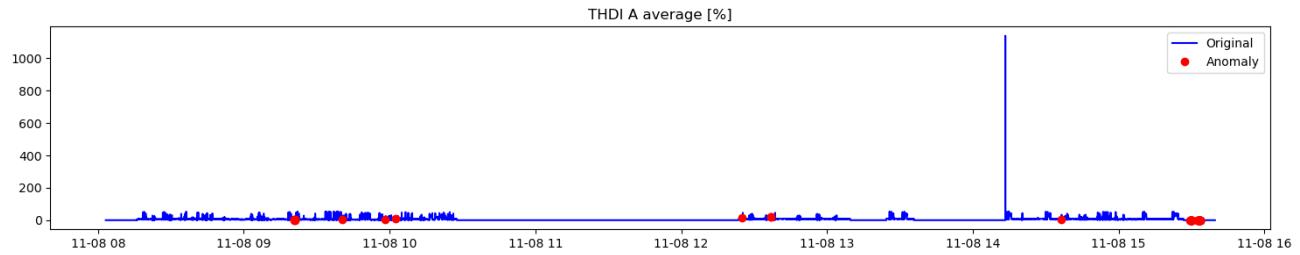
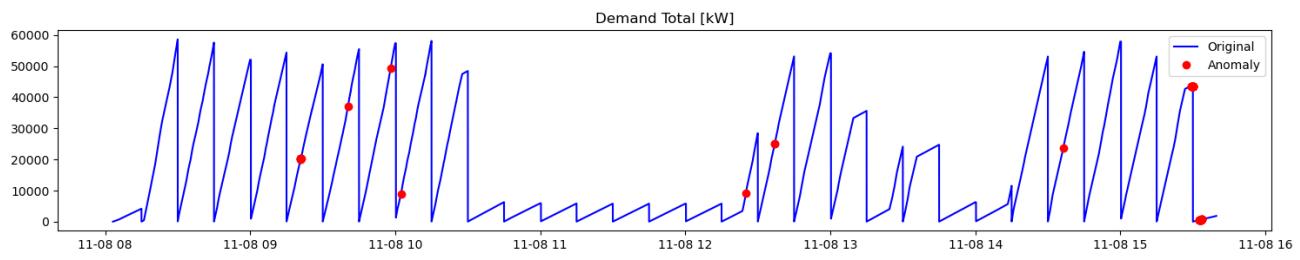


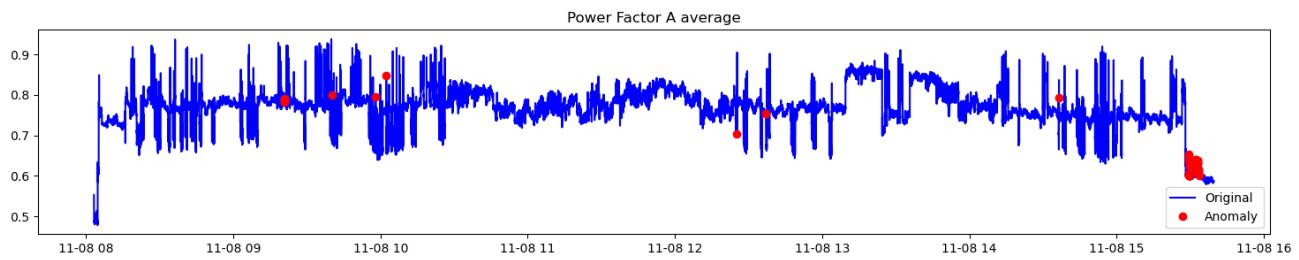
Slika 10. Drugi dan, IsolationForest(`n_estimators=50, max_samples=400, contamination= 0.001, bootstrap=True, random_state=42`).

1. Detektovana 4 od 5 pika visokih vrednosti u Active Power A average.
 2. Detektovane najnize vrednosti u Voltage A i B.
 3. Detektovane anomalije u periodu izmedju 15 i 16 h imaju interesantan oblik u Voltage C atributu.
- Period kada srednja vrednost Voltage C naglo raste i zadržava tu novu vrednost je označen kao anomaličan.

Isprobace se i `contamination 0.002` kako bi se možda detektovalo svih 5 pikova najviših vrednosti u Active Power A average.



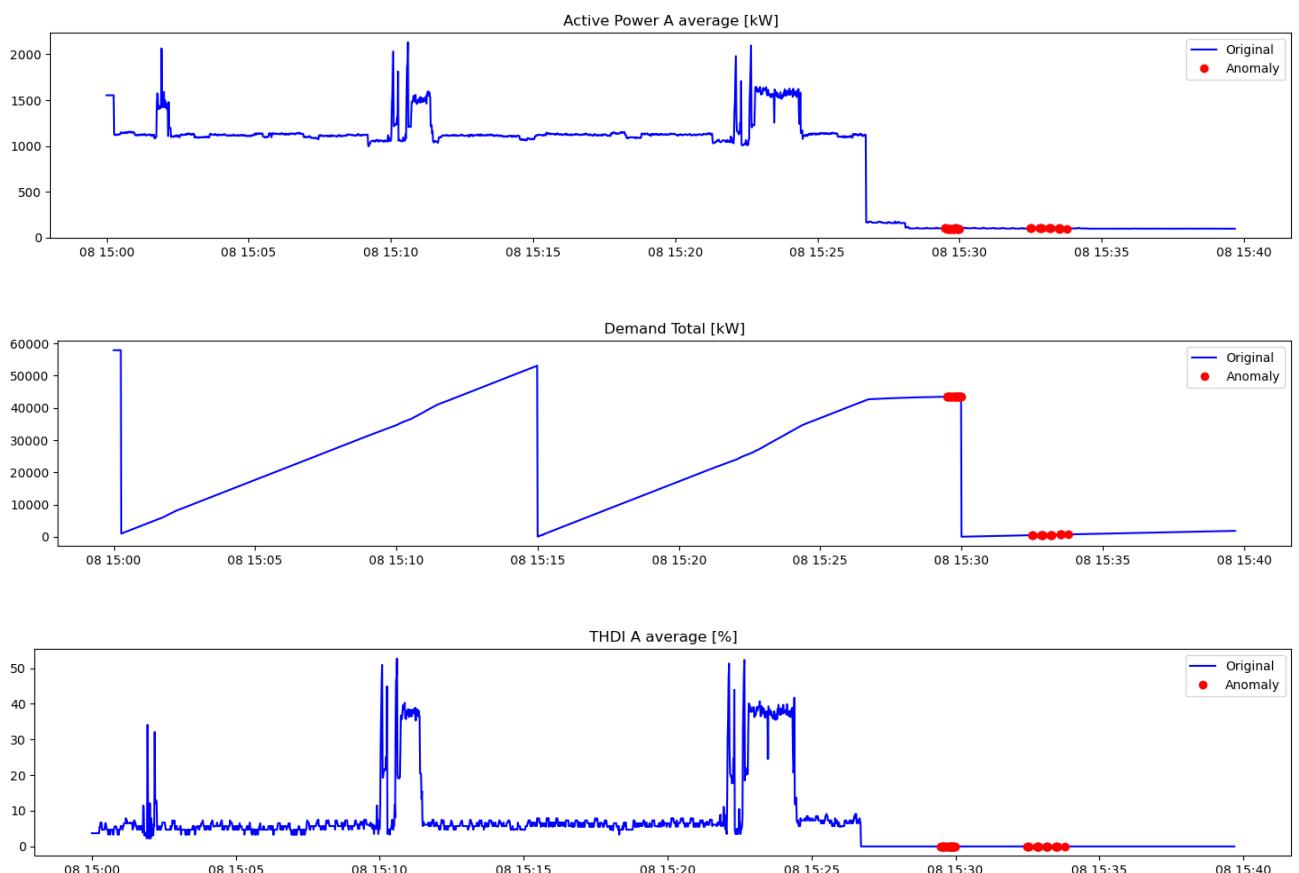


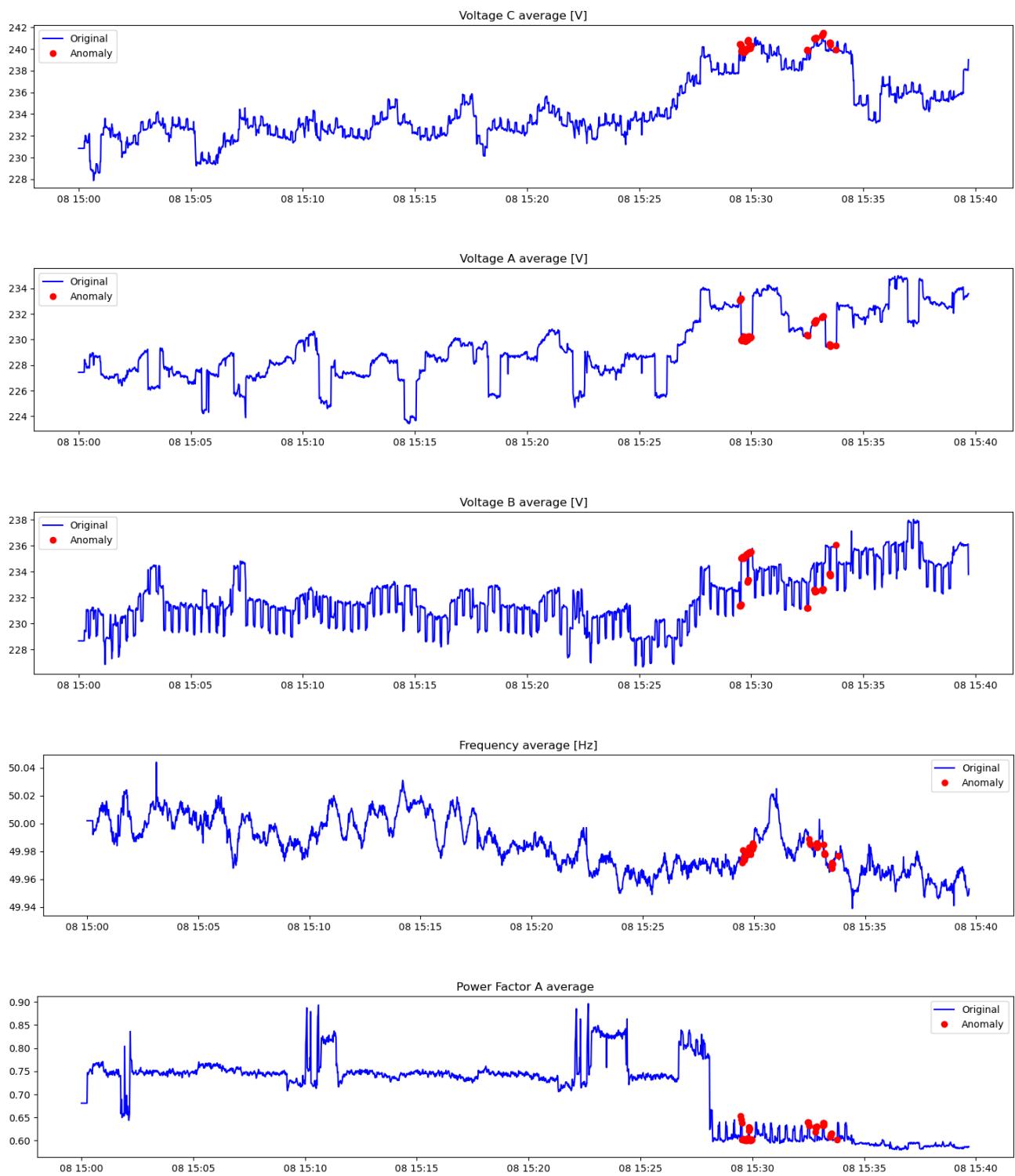


Slika 11. Drugi dan, Isolation Forest($n_{estimators}=50$, $max_samples=400$, $contamination= 0.002$, $bootstrap=True$, $random_state=42$).

1. Detektovano 5 pikova najviših vrednosti u Active Power A average.
2. Detektovan pik najniže vrednosti u Voltage A average.
3. Detektovan pik najniže vrednosti u Voltage B average.
4. Anomalije još gušće koncentrisane u periodu između 15h i 16h (konkretno u periodu kada Voltage C average ima visoku srednju vrednost u odnosu na ostatak sekvence za period između 15h i 16h).

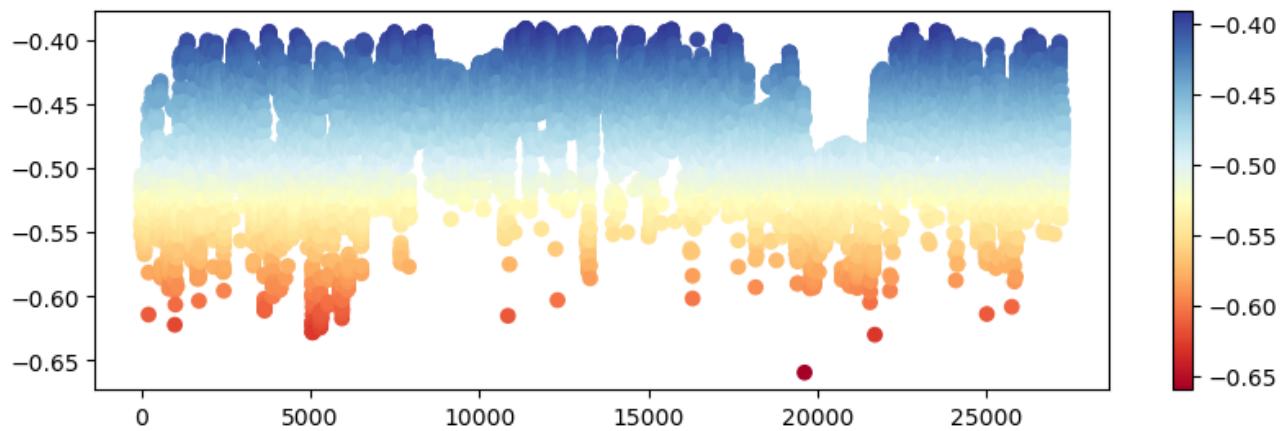
Potrebno je dodatno analizirati taj period.





Slika 12. Drugi dan, Isolation Forest (n_estimators=50, max_samples=400, contamination= 0.002, bootstrap=True, random_state=42). Period između 15h i 16h. Sumnja se da Voltage C average najviše utiče na anomalijačnost ovog perioda.

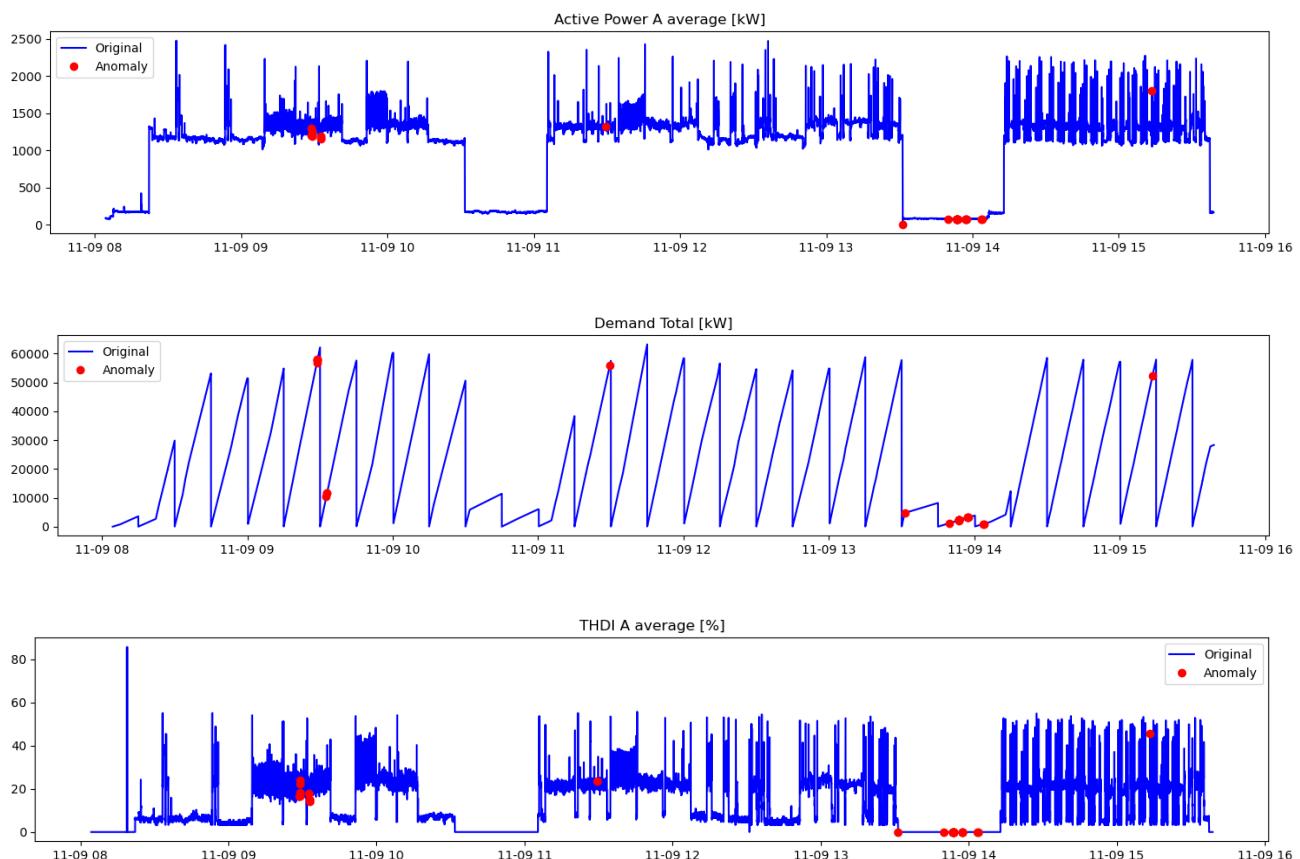
Treci dan:

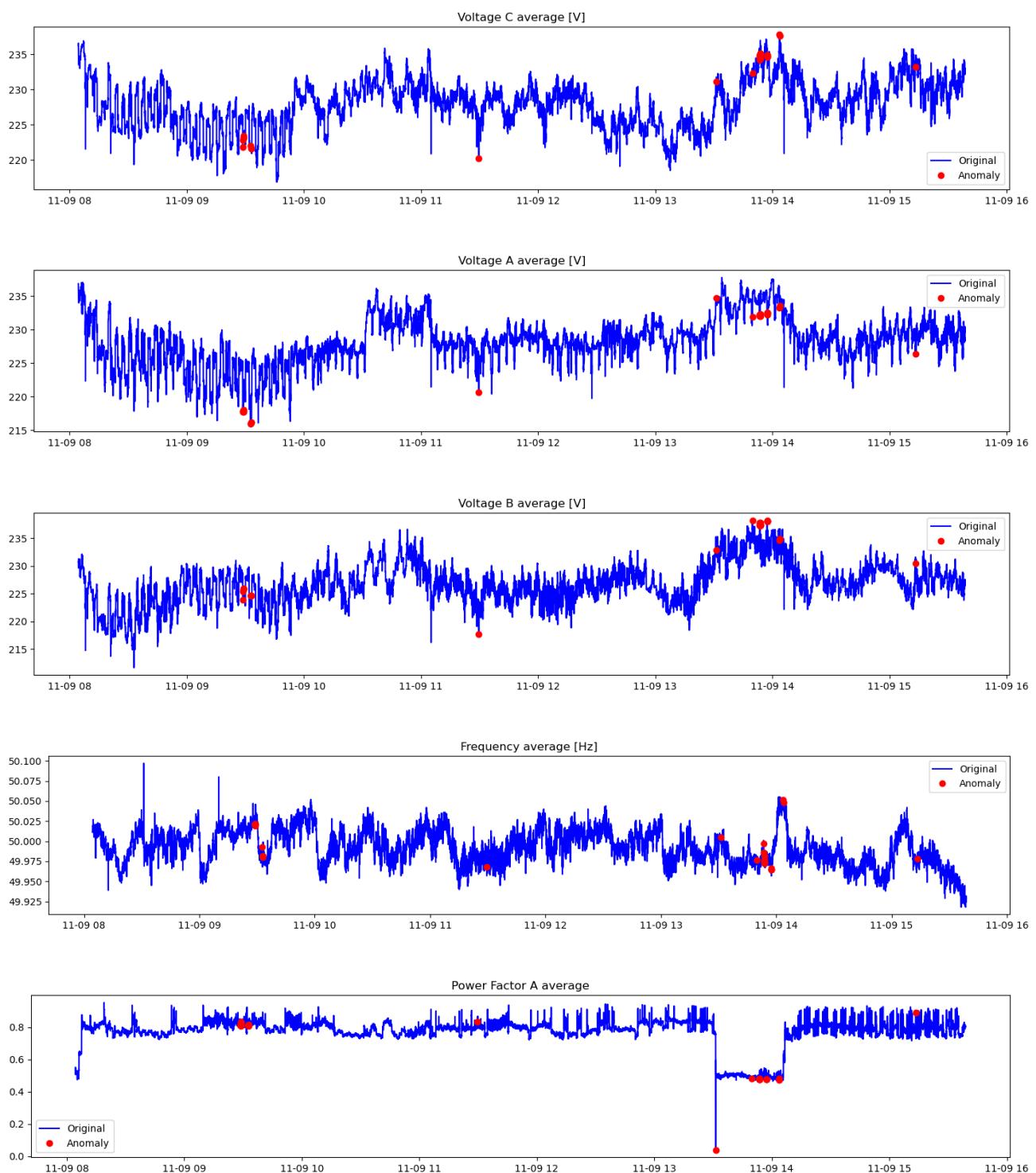


Slika 13. Score samples za treći dan.

Ovde je teško odrediti početnu kontaminaciju, pošto opservacije visoke anomaličnosti nisu toliko jasno izdvojeni. Odmah se pokreće GridSearch.

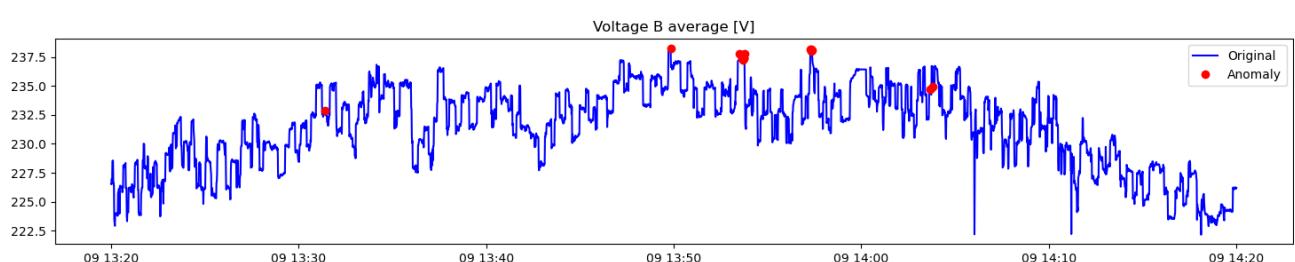
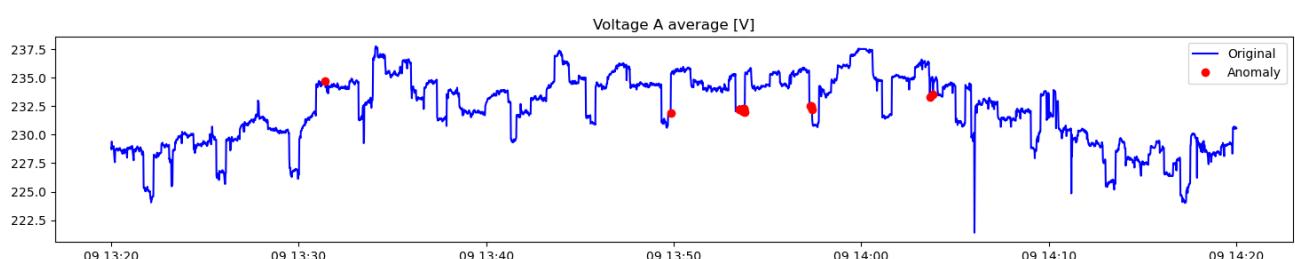
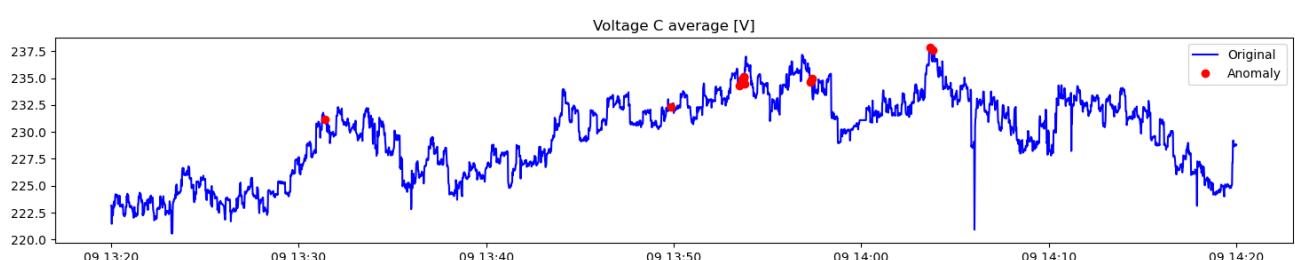
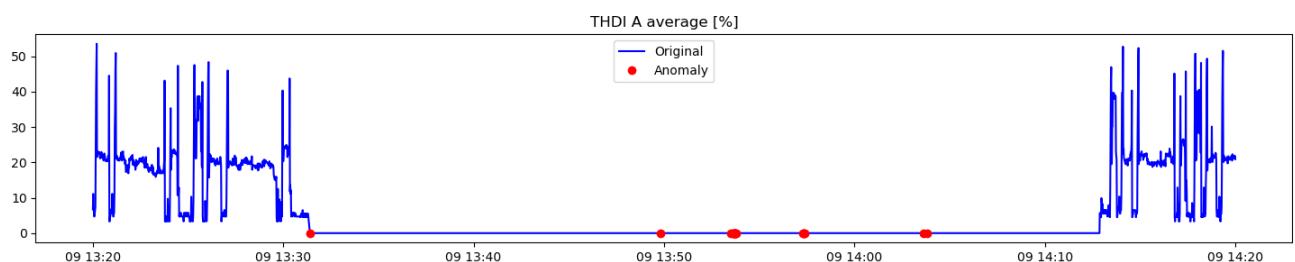
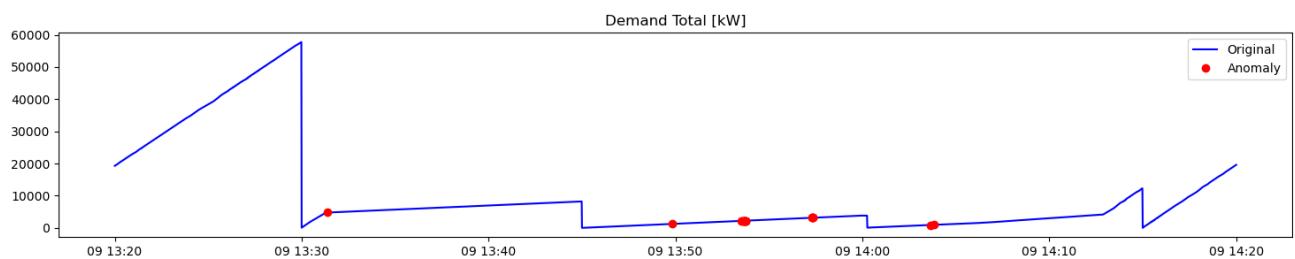
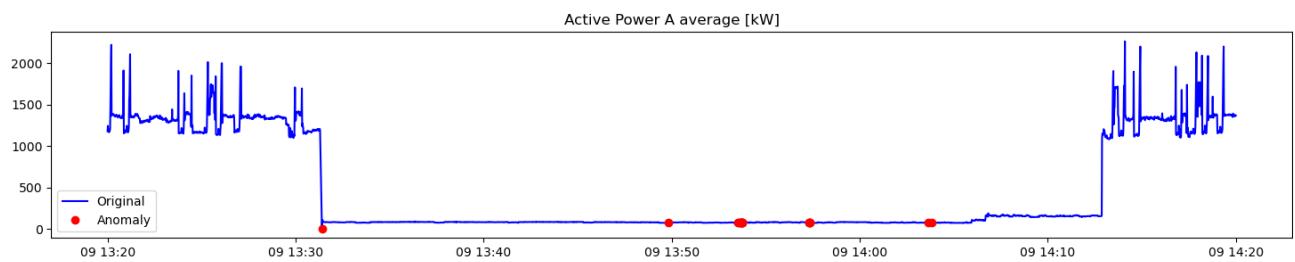
GridSearch je vratio sledeće parametre: n_estimators=200, max_samples=100, contamination= 0.001, bootstrap=False, random_state=42. Isprobane su još i contamination vrednosti 0.005 i 0.0007 u kombinaciji sa ostatkom parametara koje je vratio GridSearch. U ovom slučaju, iako kontaminacija od 0.0007 daje malo bolji silhouette score, odabrat će se 0.001 jer je detektovati više potencijalno lažnih anomalija nego ne detektovati potencijalno pravu anomaliju.

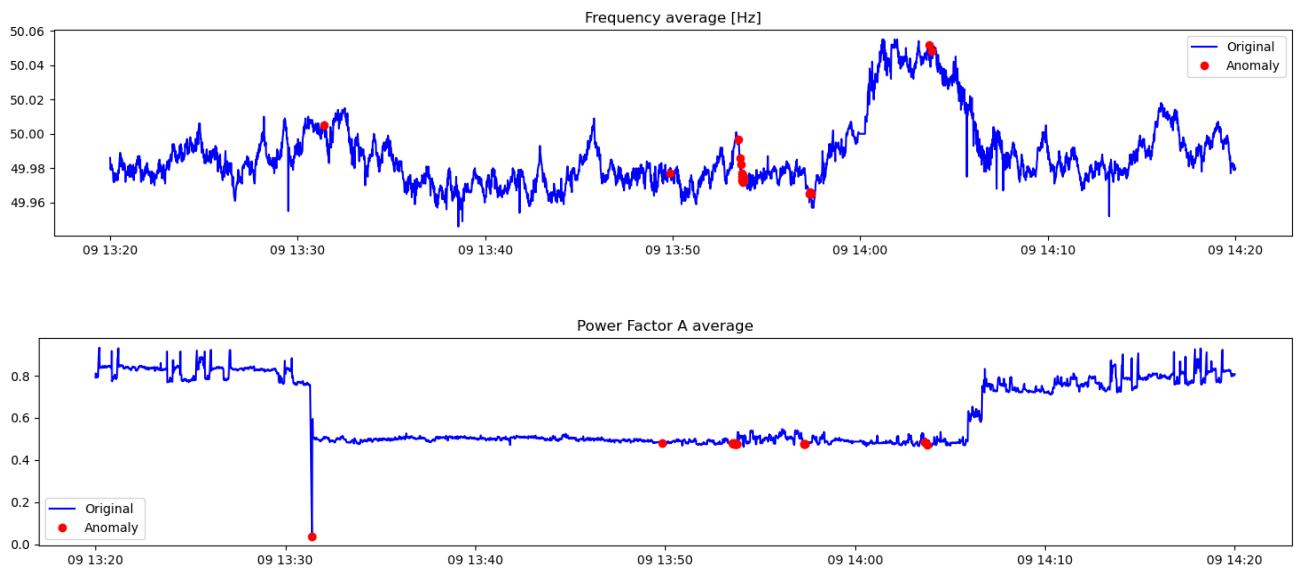




Slika 14. Treci dan, Isolation Forest(n_estimators=200, max_samples=100,contamination= 0.001, bootstrap=False, random_state=42).

Od interesa je period između 13:20 i 14:20. Treba ga dodatno analizirati.



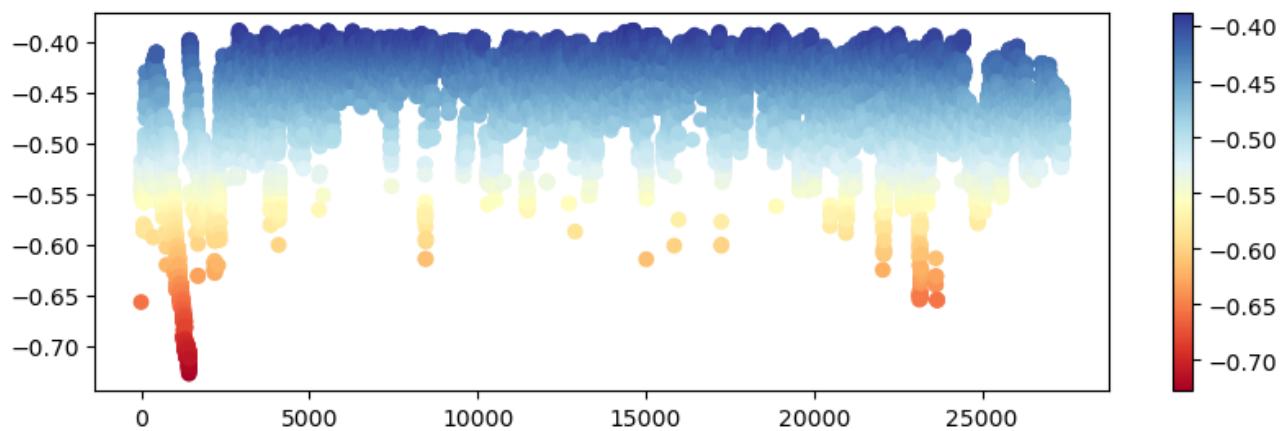


Slika 15. Treci dan (između 13:20 i 14:20) , Isolation Forest (n_estimators=200, max_samples=100,contamination= 0.001, bootstrap=False, random_state=42).

1. Pik najniže vrednosti u Power Factor A average u 13:31:24.
2. Nema nekih očiglednih pravilnosti.

Analizirajući prethodne dve slike, sumnja se da period tokom kog Power Factor A average ima nisku srednju vrednost u odnosu na ostatak tog dana dosta utiče na anomaličnost opservacija tokom trećeg dana. Ovu tvrdnju ojačava i činjenica da tokom prvog perioda niske srednje vrednosti paramtera Active Power A average nije bilo detektovanih anomalija, što nije skučaj kod drugog takvog perioda (kada je i Power Factor A average na niskoj srednjoj vrednosti).

Četvrti dan:



Slika 16. Score samples za četvrti dan.

I kod četvrtog dana će se odmah pokretati GridSearch algoritam. GridSearch je vratio parametre: n_estimators=500, max_samples=700, contamination= 0.001, bootstrap=True, random_state=42. Kao i

kod prethodnih dana, kontaminacija je bila dodatno ispitivana i odlučeno je da ona bude 0.001, budući da je vratila najviši silhouette score.

Local Outlier Factor:

Za svaki dataframe je primjenjen GridSearch algoritam za nalaženje potencijalno optimalnih parametara za LocalOutlierFactor algoritam. Parametri od koji je GridSearch birao optimalni su sledeći:

n_neighbors: [100, 120, 150, 170, 200]

contamination:['auto', 0.001, 0.01, 0.1]

algorithm:['auto', 'ball_tree', 'kd_tree', 'brute']

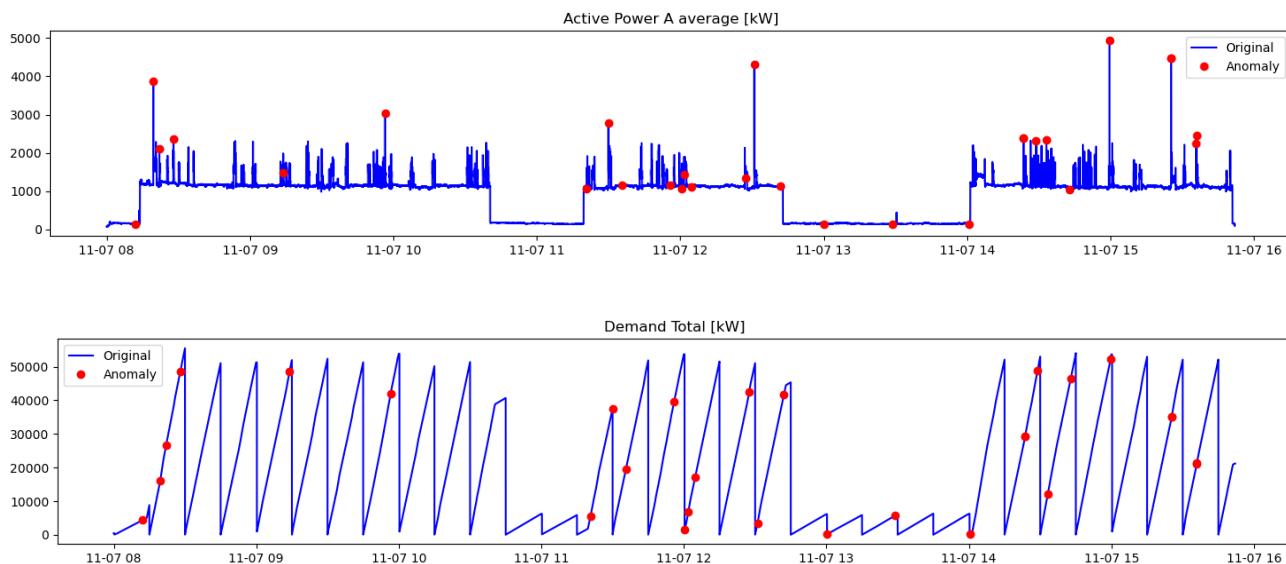
n_jobs: [-1]

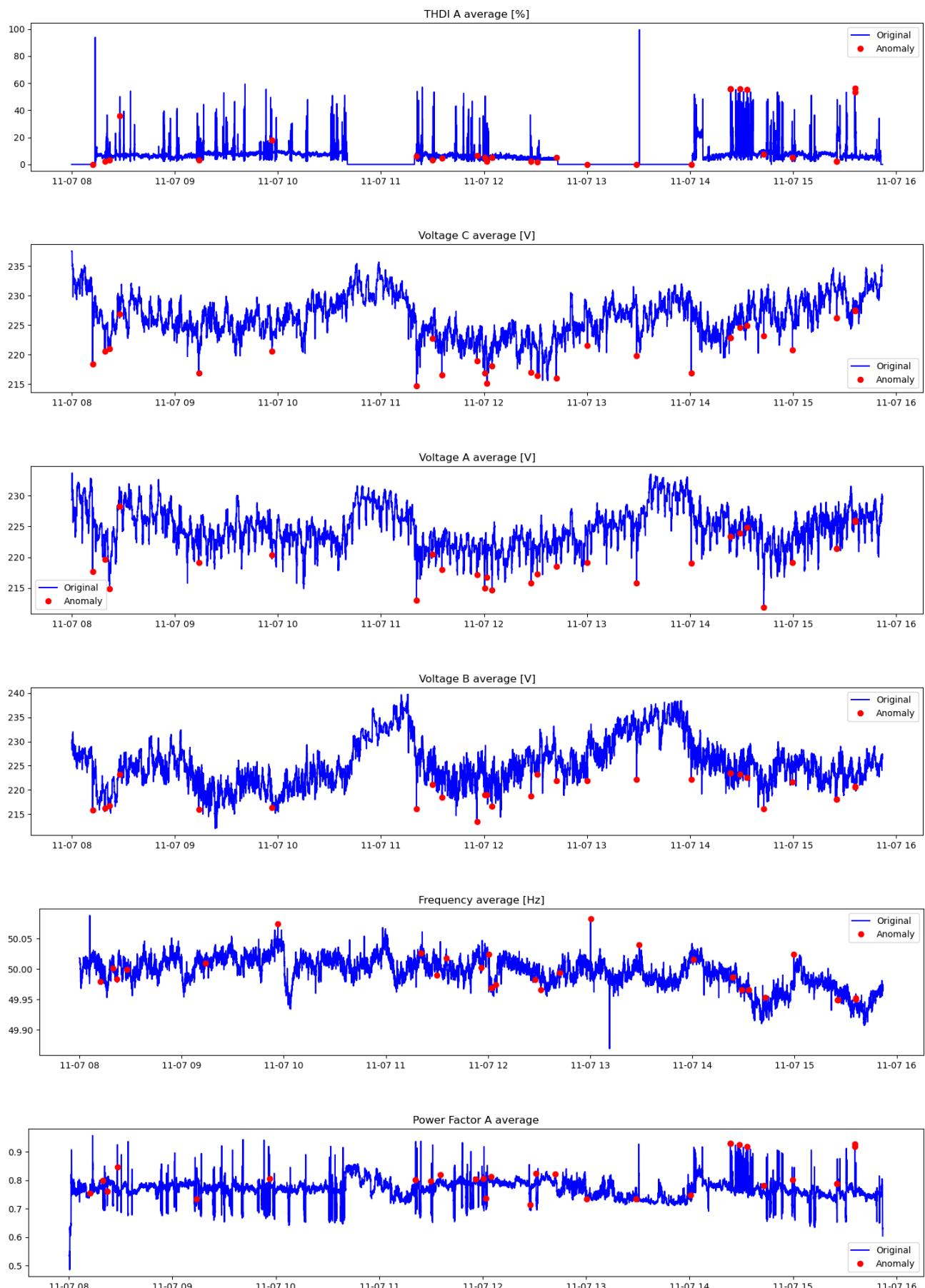
Nakon rezultata GridSearch algoritma, ukoliko bi za neki od parametra koji se ispituju bila izabrana minimalna ili maksimalna vrednost od zadatih, onda bi se taj parameter dodatno ispitivao tako što bi se pokretao algoritam LocalOutlierFactor za veću i manju vrednost tog parametra od vrednosti koju je vratio GridSearch. Na kraju se formira skup konačnih parametara IsolationForest algoritma, algoritam se pokreće, analiziraju se dobijeni rezlultati i čuvaju se u csv fajl.

Prvi dan:

Pokreće se GridSearch algoritam, koji vraća parametre:

algorithm='auto',n_neighbors=170,contamination=0.001. Kontaminacija se dodatno proveravala, i 0.001 je vratilo najviši silhouette score.



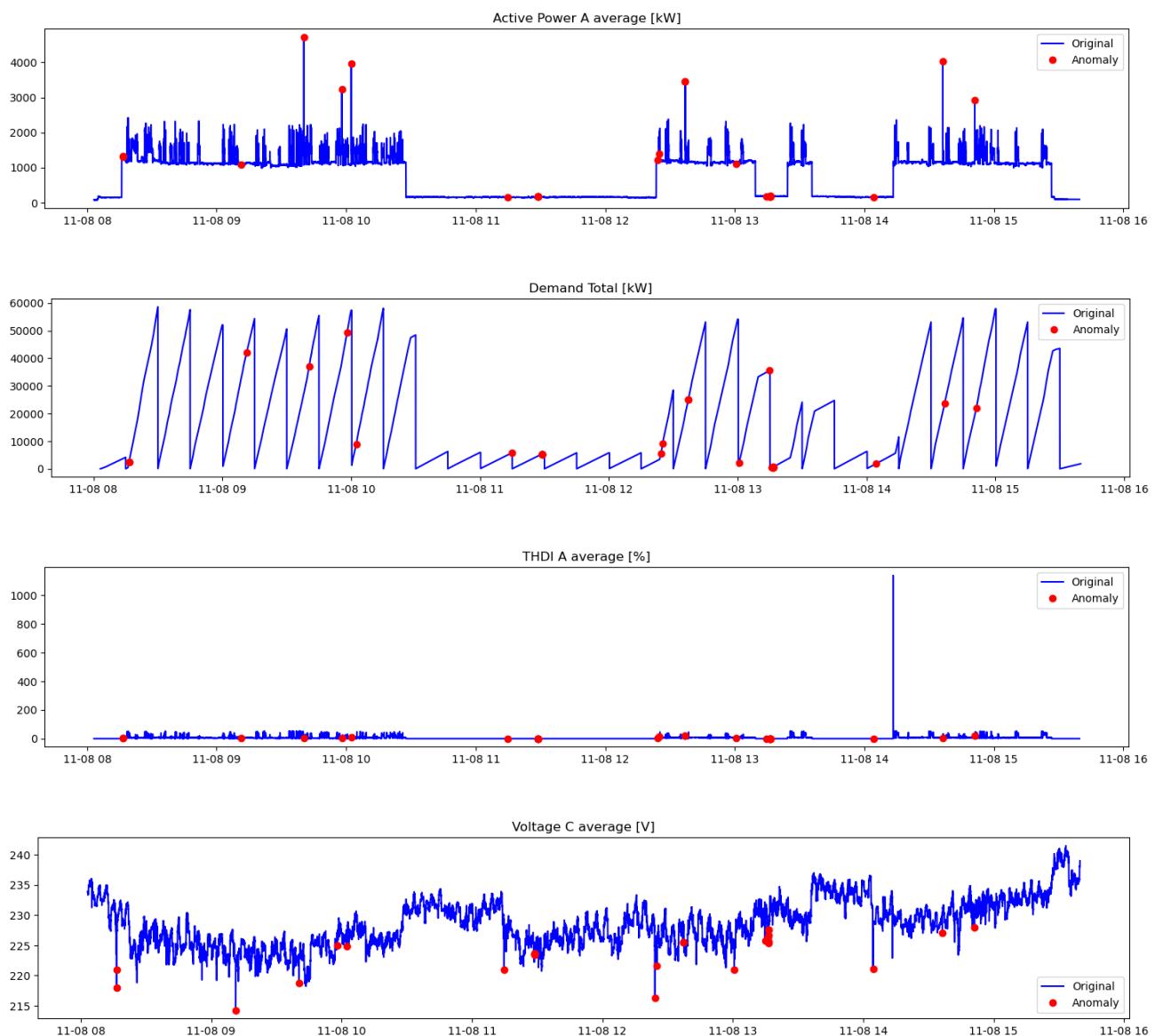


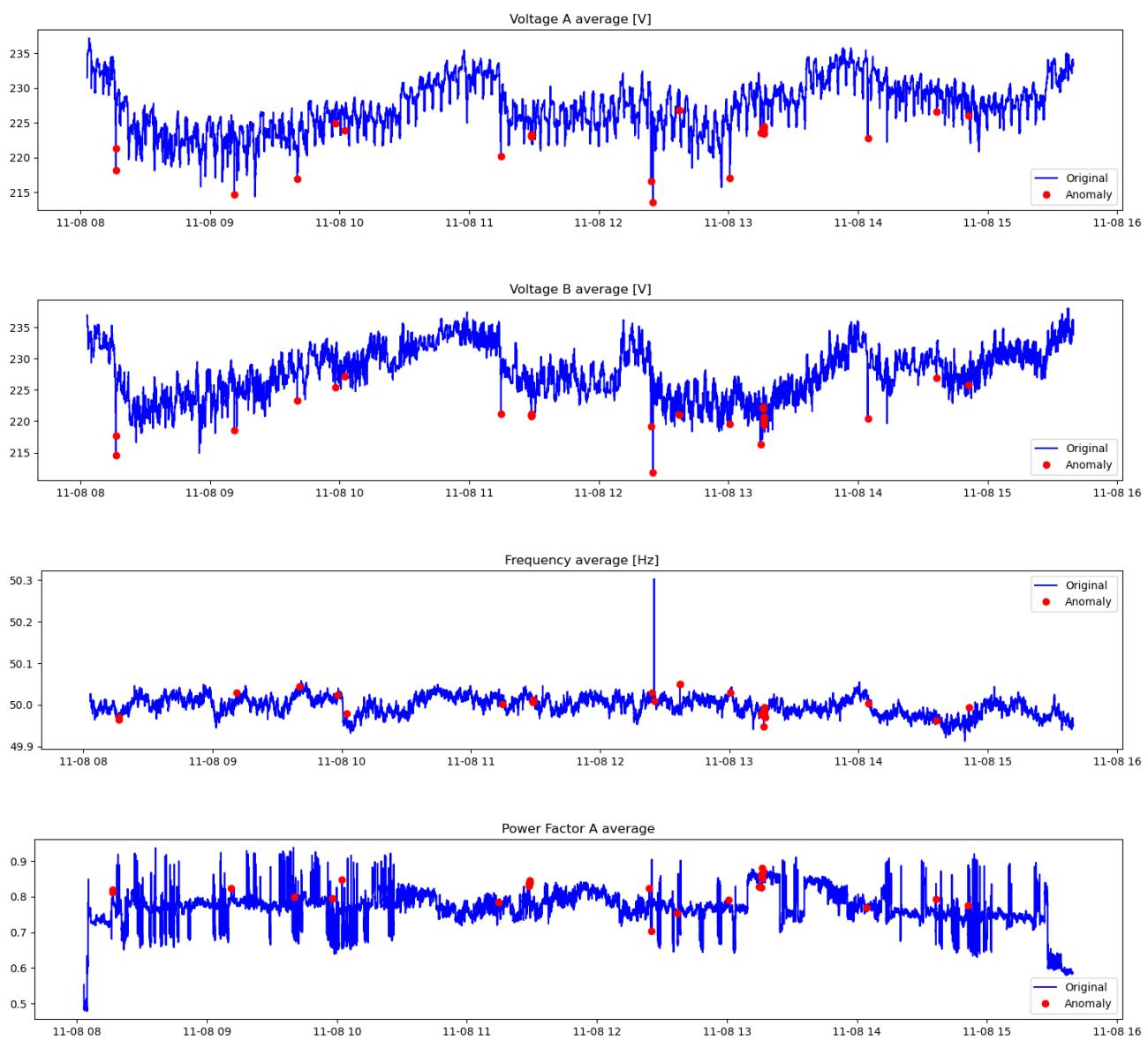
Slika 17. LocalOutlierFactor(algorithm='auto',n_neighbors=170,contamination=0.001)

1. Detektovani pikovi najviših vrednosti u Active Power A average.
 2. Detektovano puno pikova niskih vrednosti u Voltage A average, Voltage B average i Voltage C average.
 3. Pik jedne od dve najviših vrednosti u Frequency average.

Drugi dan:

GridSearch algoritam je vratio parametre: algorithm='auto',n_neighbors=100,contamination=0.001. Dodatnim ispitivanjem kontaminacije je utvrđeno da 0.0007 daje malo viši silhouette score, međutim ipak je odabранo 0.001.



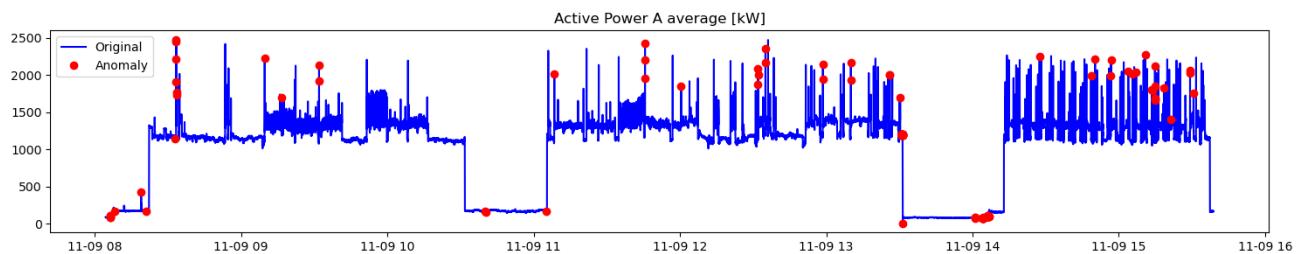


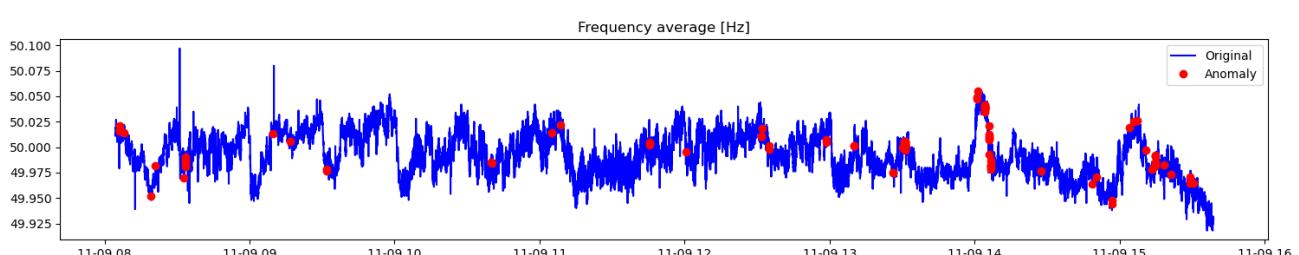
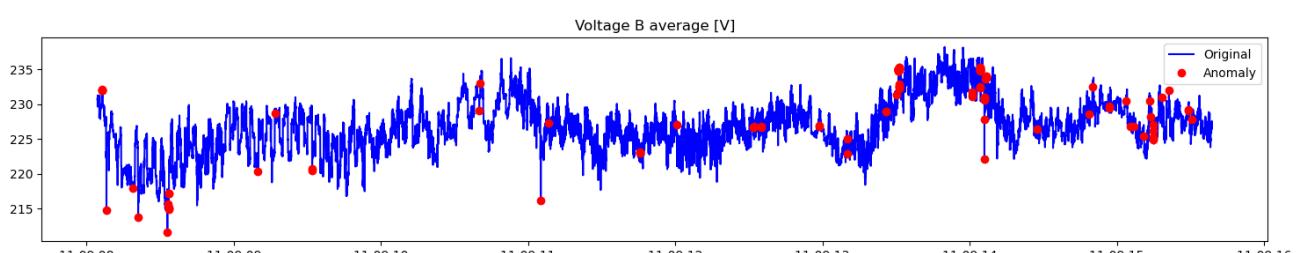
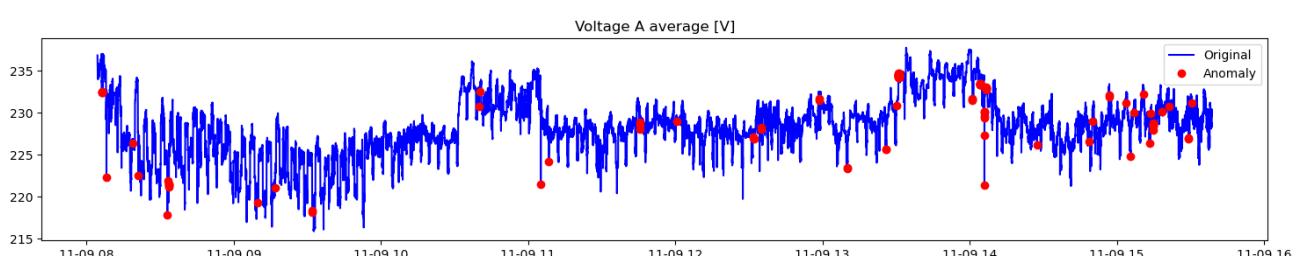
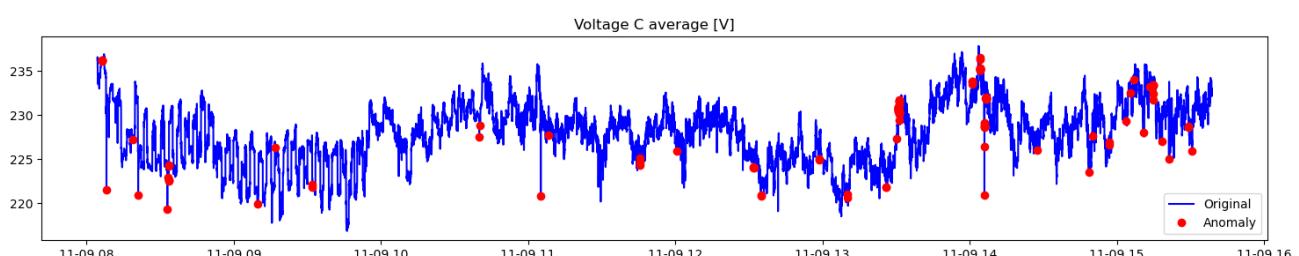
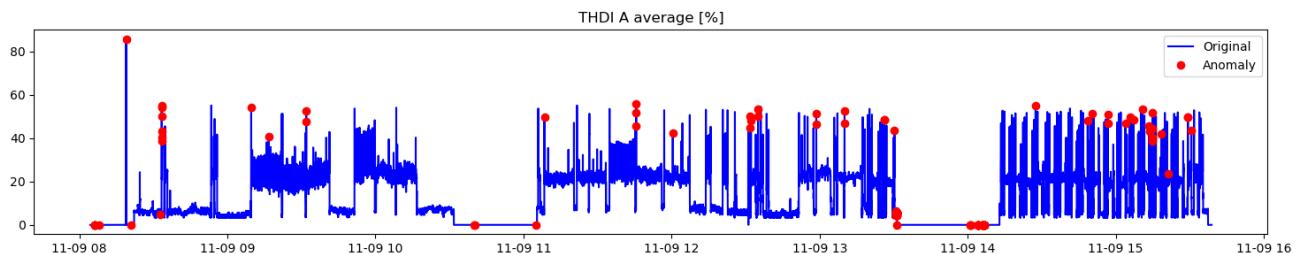
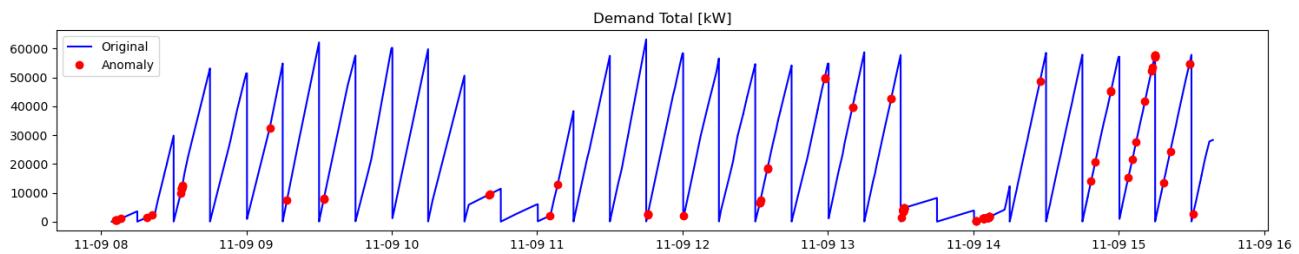
Slika 18. LocalOutlierFactor(algorithm='auto',n_neighbors=100,contamination=0.001).

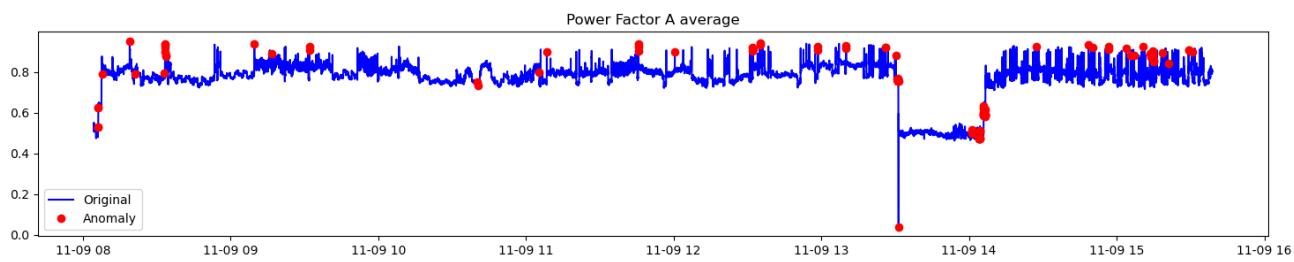
1. Detektovani pikovi najviših vrednosti u Active Power A average.
2. Detektovani pikovi niskih vrednosti u Voltage A avearge, Voltage B average i Voltage C average.

Treći dan:

GridSearch algoritam je vratio: algorithm='auto',n_neighbors=100,contamination=0.001. Dodatnim ispitivanjem kontaminacije je utvrđeno da 0.004 daje najviši silhouette score, pa je ona i odabrana.





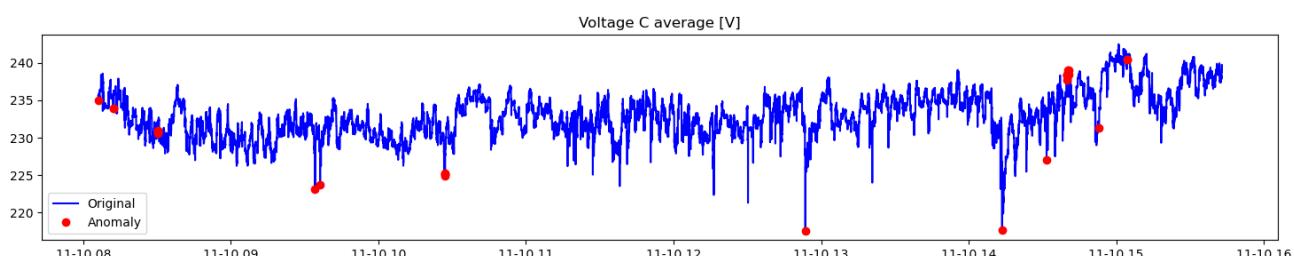
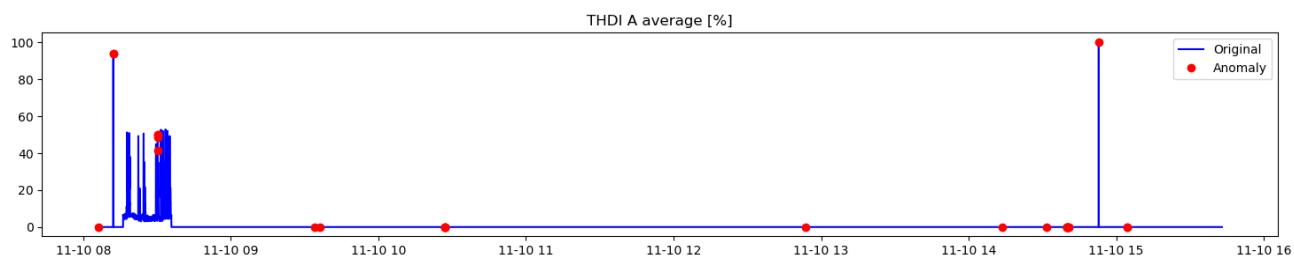
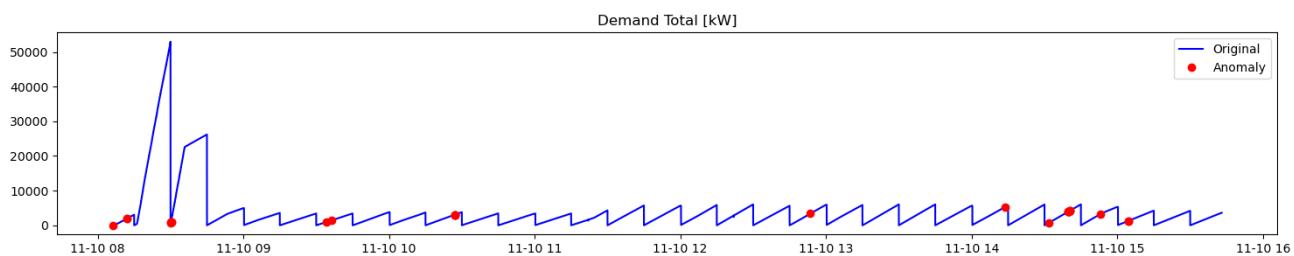
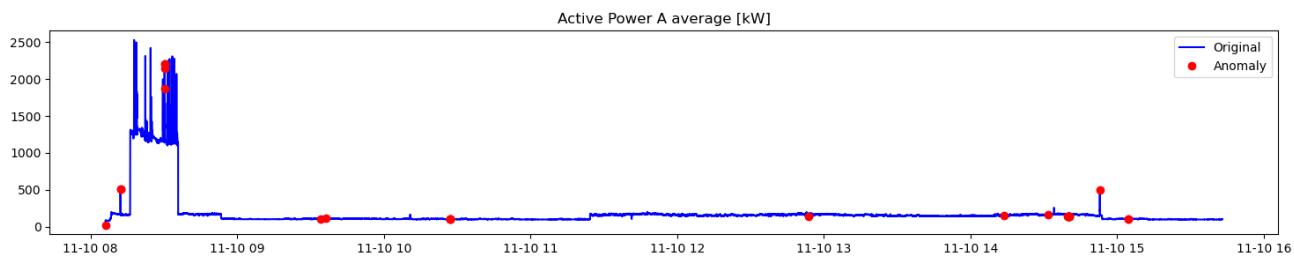


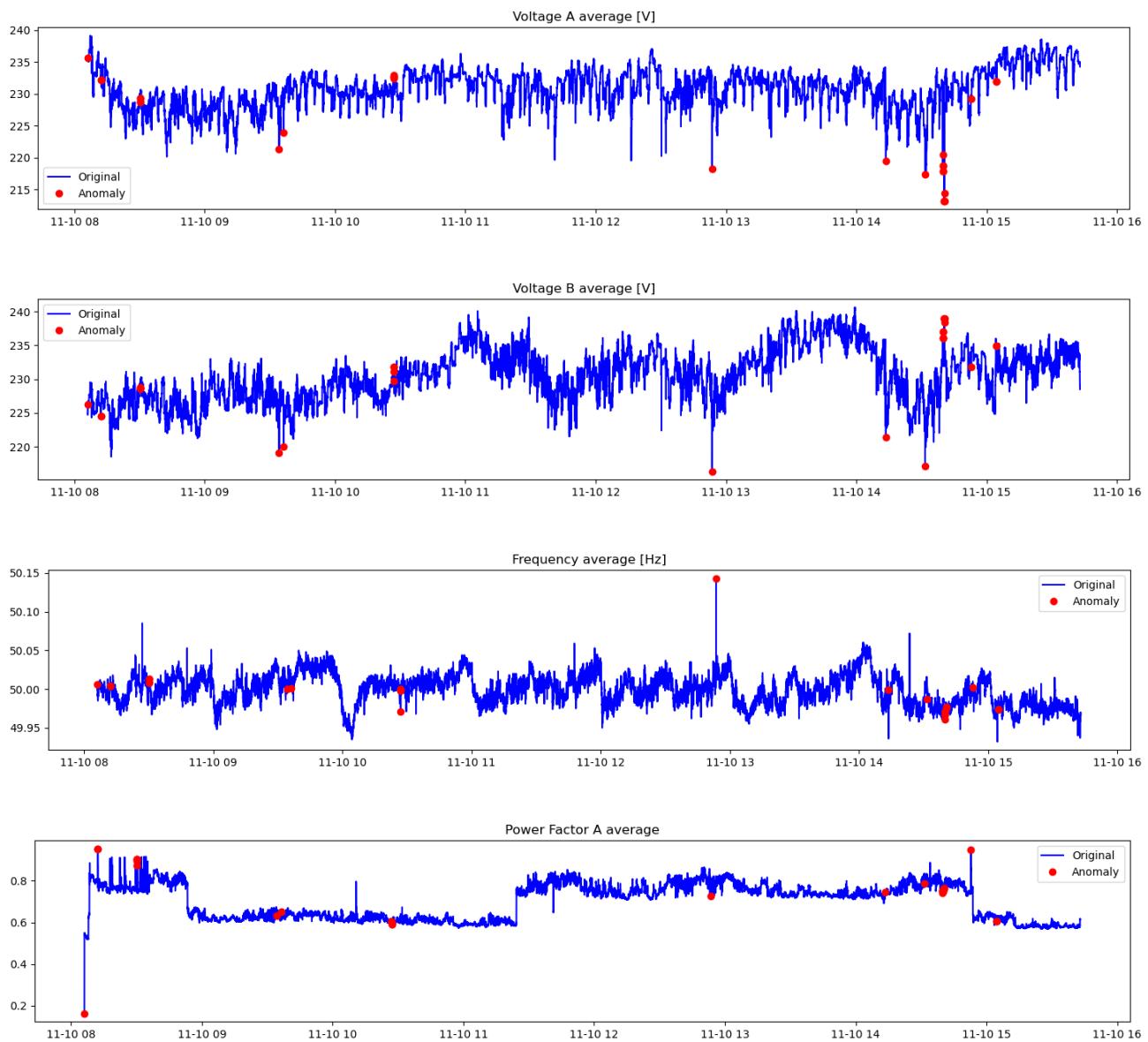
Slika 19. LocalOutlierFactor(algorithm='auto',n_neighbors=100,contamination=0.004).

Za razliku od IsolationForest, ovde period između 13:20 i 14:20 nije toliko izdvojen. Pik u 13:31:24 je i ovde detektovan.

Četvrti dan:

GridSearch algoritam je vratio parametre: algorithm='auto',n_neighbors=100,contamination=0.001. Kontaminacija je i ovde dodatno ispitivana i 0.001 daje najviši silhouette score.





Slika 20. LocalOutlierFactor(algorithm='auto',n_neighbors=100,contamination=0.001).

1. Detektovana dva najviša pika u THDI A average.
2. Detektovan pik najviše vrednosti u Frequency average.
3. Detektovan pik najniže i dva pika najviših vrednosti u Power Factor A average
4. Za razliku kod IsolationForest, ovde period između 8 i 9 časove ne utiče vidno na ishod algoritma.