Favorita store sales demand forecasting

-Prognoziranje vremenske serije-

ZAJIĆ NEMANJA 2022 FON

KATEDRA ZA ODLUČIVANJE

1 UVOD

Otpad hrane je oroman problem. Globalno, ukupna vrednost bačene hrane se procenjuje na £2.9 biliona godišnje. To je dovoljno da se nahrani svaki gladan čovek na svetu dva puta. Sa očekivanom svetskom populacijom od 9.7 milijardi u 2050, ne možemo da priuštimo da bacimo korisne resurse ili hranu.

Supermarketi su delom odgovorni za katastrofu globalnog otpada hrane. Procenjeno je da lanci supermarketa godišnje bacaju hranu u vrednosti od £230 miliona. Za supermarkete, tačnije predviđanje može da smanji otpad hrane povezan sa prekomerim skladištenjem robe i poboljša zadovoljsto kupca

Za prognozu vremenske serije koriste se podaci Corporación Favorita, velikog lanca supermarketa stacioniranom u Ekvadoru. Specifično, kreiraće se modeli koji tačno predviđaju prodaju različitih jedinica familije proizvoda Favorita prodavnica.

Predviđaju se 33 familije proizvoda u 54 prodavnice sa horizontom od 16 dana, potencijalno 1,782 različitih modela. Postoje više načina modelovanja problema vremenskih serija pomoću Box-Jenxins ARIMA metodologije, Exponential smoothing, Neuronskih mrežea (RNN, GRU, CNN, LSTM), biblioteka kao što su Darts i FbProphet i ML pristupom tretirajući vremensku seriju kao regresioni problem.

Zbog teškoće u modelovanju potencijano velog broja modela odbacićemo ARIMA i NN pristup. Dok ćemo se osloniti na ML metodologiju i koristiti neke od raspoloživih biblioteka.

2 OPIS I RAZUMEVANJE PODATAKA

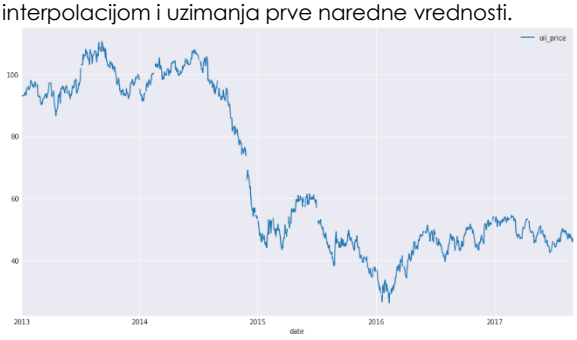
Podaci su preuzeti sa sajta kaggle, Store Sales - Time Series Forecasting takmičenja. Oni sačinjavaju 6 tabela: praznici u Ekvadoru, cene nafte jer je Ekvador naftno zavisna zemlja, opis prodavnica, transakcije prodavnica i tabele train i test u kojima se nalaze izlazne zavisne varijable. Podaci su u vremenskom rasponu od 1.1.2013 do 31.8.2017. Horizont predviđanja je od 16.8.2017-31.8.2017

2.1 OPIS ATRIBUTA

Train dataset čine podaci u rasponu 1.1.2013-15.8.2017, u njemu se nalaze atributi koliko je jedinica familije proizvoda prodato u kojoj prodavnici i koliko je jedinica proizvoda bilo na promociji. Test čini horizont predviđanja bez jedinica prodaje, ali sa promocijom, testiranje prognoze će se obaviti prosleđivanjem rezultata prodaje na kaggle. Dataset sa cenom nafte je dat nad celim vremenskim rasponom, ti podaci čine takozvani **data leakage**! Dataset sa transakcijama je dat u rasponu do početka horizonta predviđanja. Što znači da ako bi smo koristili te podatke u predviđanju morali bi i njih da predvidimo. Oni su takođe **data leakage!** U datasetu sa datumima praznika su opisani tipovi praznika, da li je događaj, praznik ili produženi vikend. Da li postoji lag (npr Nekoliko dana pre Božića) i koja je vrsta praznika , da li je nacionalni ili lokalni. U datasetu opisa prodavnica dato je kog je tipa prodavnica, kom klasteru pripada i njena lokacija.

2.2 NEDOSTAJUĆE VREDNOSTI

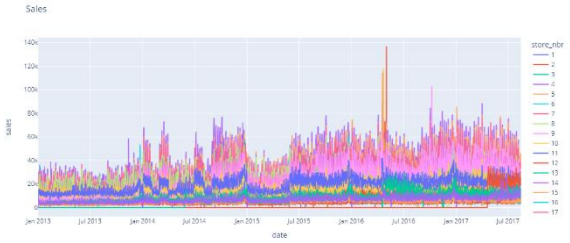
U datasetu nafte ima 43 nedostajuće vrednosti. Popunjavanje nedostajućih vrednosti se izvršilo interpolacijom i uzimanja prve naredne vrednosti.



Nafta-sa nedostajućim vrednostima

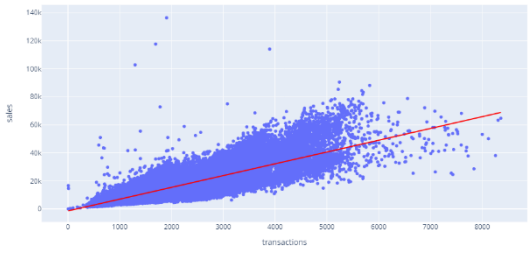
Nakon imputacije vrednosti interpolacijom testirana je stacionarnost serije nafte ADF testom i utvrđena je nestacionarnost. Nakon pretvaranja u stacionarnu prvim stepenom diferenciranja proverena je njena zavisnost u odnosu na lag paricalnom autokorelacionom funkcijom. Utvrđeno je da je ona tip serije Nasumičnog hoda (Random Walk) i jedina odgovarajuća predikcija za tu seriju je **Naive forcasting.**

2.3 VIZUELIZACIJA I ZAVISNOST U PODACIMA



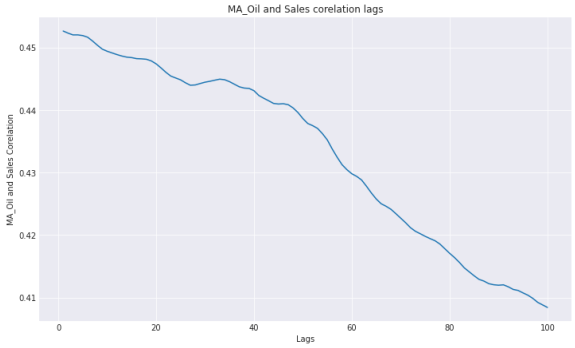
TS ukupne prodaje prodavnica

Primećujemo dosta nestabilnosti u prodaji nakon čega se ona stabilizovala u junu 2015. Zbog toga će podaci biti uzeti nakon tog datuma za trening našeg modela.
Dok funkcije tansakcija nemaju tu nestabilnost i na osnovu toga zaključujemo probleme u nabavci robe. Nakon nekog vremena neke prodavnice su bile privremeno zatvorene, dok su se neke kasnije otvorile. Takođe prodavnice su zatvorene prvog dana u novoj godini. Ono što isto možemo da primetimo je dolazak više kupaca pred Božič. Dok na prvom grafu prodaje vidimo impuls koji je nastao nakon **velikog zemljotresa** koji je pogodio Ekvador 18.4.2016.



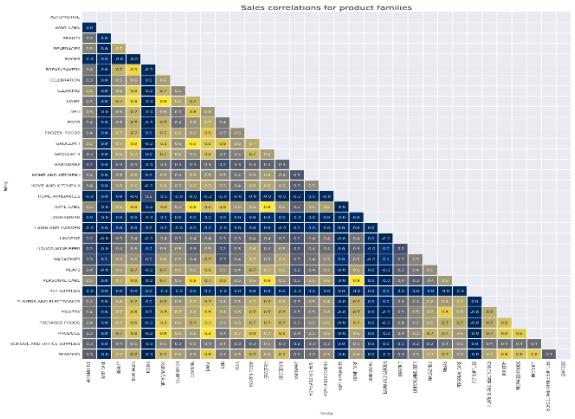
Odnos Transkakcije i Ukupne prodaje

Testirana je prosečna zavisnost između prodaje familije proizvoda i cene laga nafte. Za upoređivanje je uzeta i nedeljna cena Moving Avg nafte i dnevna cena.



Korelacija MA laga nafte sa prodajom

Na osnovu funkcije korelacije laga nafte i prodaje uzeta su 3 MA laga da pomognu u predviđanju modela. Razlog je što primećujemo dosta ispupčenja u funkciji korelacije cene nafte i prodaje.



Corelacija proizvoda

*Sa grafika primećujemo dosta međusobno korelisanih familije proizvoda, što ovo čini **multivarijantnu vremensku seriju**. Familije koje nisu korelisane su BABY CARE, BOOKS, HOME APPLIANCES, LADIESWEAR, LAWN AND GARDEN i PET SUPPLIES. Poznata je činjenica da su robe iz grupe FMCG međusobno korelisane.*

3 PRIPREMA PODATAKA

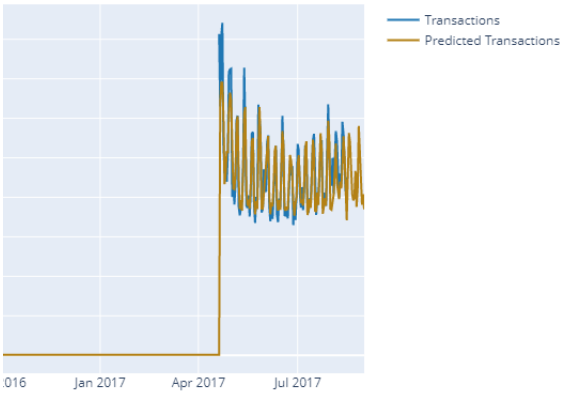
TS cene nafte

Sa grafika se vidi izled funkcije nafte nakon imputacije interpolacijom.

Zbog curenja podataka iz budućnosti morali smo napraviti modele za predviđanje transakcija ako zelimo da nam pomognu u predviđanju prodaje. Uzet je LGBMRegressor i koristili smo DirRec metodologiju pomoću RegressorChain-a. Ulazni podaci su nedeljne transakcije sa horizontom velilčine 16. Parametri su optimizovani kros validacijom sa 5 odsečaka. Za svaku prodavnicu je napravljen poseban model.

Prodavnica	RMSLE
49	0.2559
50	0.2718
51	0.2486
52	6.3804
53	0.3265
54	0.2385

Predvidjanje broja transakcija u prodavnici



Transakcije prodavnice 52

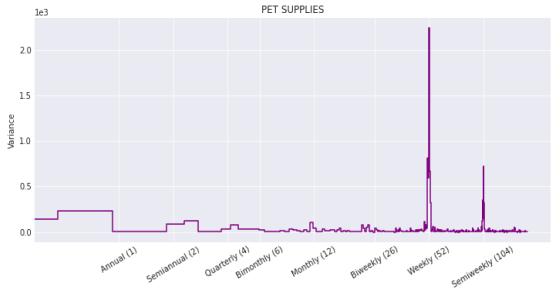
Sa grafika vidimo zašto je malo veća greška, jer se ta prodavnica najkasnije otvorila, dok ona nije radila model za to vreme predviđa 6 transakcija, kasnije su ti podaci izbačeni.

parametri	vrednosti
max_depth	3
min_child_samples	10
n_estimators	70
reg_lambda,	1

Najbolji parametri modela transakcije prodavnice 52

Zbog ML pristupa u predviđanju, morali smo da napravimo sezonske indikatore i indikatore praznika. Većina familija proizvoda ima nedeljnu sezonalnost, gde najviše posetilaca ima subotom, a najmanje četvrtkom.

Testiranje sezonalnosti svakog proizvoda je izvršeno Periodogramom.



Periodogram za PET SUPPLIES

Izuzetak su školski proizvodi, koji se ponavljaju dva puta godišnje i traju mart-april i avgust-septembar. Na GROCERY I, HOME CARE, PERSONAL CARE prodaju je uticao zemljotres, par dana nakon pojavljuju se impulsi. Dok na prodaju FROZEN FOODS i LIQUOR utiče blizina Božića.

Za predviđanje Multivarijantne vremenske serije dodati su i rezultati predviđanja korelativnih proizvoda. Model LGBMRegressor je uzet sa navedenim parametrima.

parametri	vrednosti
n_estimators	150
min_child_samples	10
max_depth	7
reg_lambda	1
sub_sample	0.8

Ulazni podaci su indikatori koji smo napravili do tada. Zbog pojavljivanja problema multikorelacije među novo dodatim atributima, dodat je PCA pipeline u finalni model sa 95% varijabiliteta.

Dodatni atributi za ne korelisane proizvode su uzeti pomoću biblioteke UpGINI, koja ih nalazi pomoću ključa Zemlje i datuma i automatski im testira shap vrednost i koliko je poboljšana tačnost.

Za PET SUPPLIES proizvode:

atributi	shap
f_holiday_code_c0eb7712	0.046270
f_c2c_fraud_score_5028232e	0.021629
f_cbpol_pca_5_b4583d1b	0.021255
f_cbpol_pca_9_0be25bd8	0.016847

Baseline MSE	Enriched MSE	uplift
0.025239	0.025699	-0.00046

Kao što vidimo dodavanjem novih atributa za proizvod PET SUPPLIES neznatno je smanjena tačnost predviđanja. Napravljeni su posebni obogaćeni datasetovi za svaku ne korelisanu familiju proizvoda.

4 MODELOVANJE

Napravljena su dva prilagođena modela, jedan hibridni koji odvaja ne stacionarne i stacionarne parove prodavnica-proizvod i jedan sa jednim vrstom modela. Hibridni model sačinjavaju boosted tip modela XGBMRegressor i Ridge, dok je drugi samo LGBMRegressor. Oba modela imaju različite pipeline-ove u odnosu na to da li je proivod korelisan sa drugim ili ne. Zbog velikog broja potencijalnih modela i mogućeg dužeg trajanja treninga uvedeno je i snimanje prilagođenog modela kada su naučeni svi modeli iz jedne prodavnice. Korišćena je i biblioteka fbprophet u modelovanju.

4.1 GRESKA MODELA

Za evaluaciju modela je korišćena RMSLE greška.

$$L(y,\hat{y}) = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}(\log(y_i+1) - \log(\hat{y}+1))^2}$$

RMSLE samo gleda relativnu razliku između stvarne i prediktivne vrednosti, ili drugim rečima, samo brine o procentualnoj razlici između njih.

Ovo znači da će RMSLE tretirati greške isto kada su prediktivne i istinite vrednosti male i kada su velike. Ona penalizuje više kada podbacimo nego kada prebacimo, ubaca asimetriju u krivi greške.

Zbog nemogućnosti ubacivanja RMSLE greške u prophet model, njegova greška na trening data setu biće iražena u MAE,RMSLE,MAPE grešci dok će u test setu biće izražena u njoj.

4.2 TRENIRANJE MODELA

Pre treninga modela primenjena je metoda **zero forecasting.** Naime tokom vremena neke prodavnice su prestale da prodaju određen proizvod, dok ga neke nisu uopšte prodavale. Zato je uveden princip da ako zadnje **dve nedelje** prodavnica nije prodala ni jednu vrstu određenog proizvoda, neće je prodati ni za narednih 16 dana. Tom metodom uspeli smo da smanjimo broj modela za 132, dakle ostalo nam je 1,650 modela za trening.

Treniranje prilagođenih modela je optimizovano tražeći najbolje parametre GridSearchCV-om i u njemu su uvedeni različiti pipeline-ovi u zavisnoti od proizvoda. Za korelacione proizvode uveden dodatan dodatan korak PCA transformacije kolona korelativnih proivoda tako da se zadržava 95% varijablilitеta. Korišćenje FbProphet biblioteke je bilo malo zahtevnije zbog RAM memorije. Pokušan je i eksperiment paralelnog učenja hibridnog modela sa dodatnim parametrima. Naime zbog treninga na rezidualima nije bilo moguće staviti RMSLE grešku, već je stavljena RMSE greška, a evaluacija na test setu biće iražena RSMLE greškom. Isto i u FbProphet-u nije moguće ubaciti prilagođenu evaluaciju RMSLE, zbog ugrađenih grešaka(MAE, MSE, RMSE , MAPE), dok će greška na testsetu biti isto izražena u RMSLE.

Paramertri XGBMRegressor-a u hibridnom modelu bez paralelizacije.

n_estimators	reg_lambda	min_child_samples	max_depth
100	2	5	7
140	1	10	10
170	0	15	15
200		20	

Trening 1,650 modela je trajao ~10h.

Trening 1,650 modela paralelnim pristupom je trajao ~8h! U ovom drugom hibridnom modelu su dodati novi parametri.

Parametri LGBMRegressor-a u drugom prilagođenom modelu.

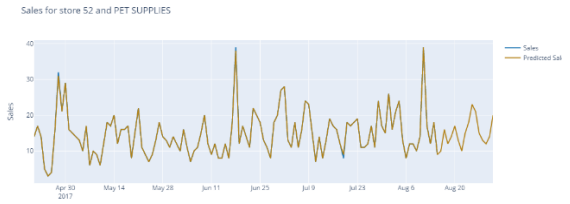
n_estimators	reg_lambda	min_child_samples	max_depth
100	2	5	7
140	1	10	10
170	0	15	15
200		20	

Trening 1,650 modela je trajao ~9h.

Već je napomenuto da je metrika na trening data setu izražena u RMSE-u, dok je na test-u u RMSLE-u. Zbog ovoga je malo teže proceniti da li je došlo do eventualnog pretreniranja.

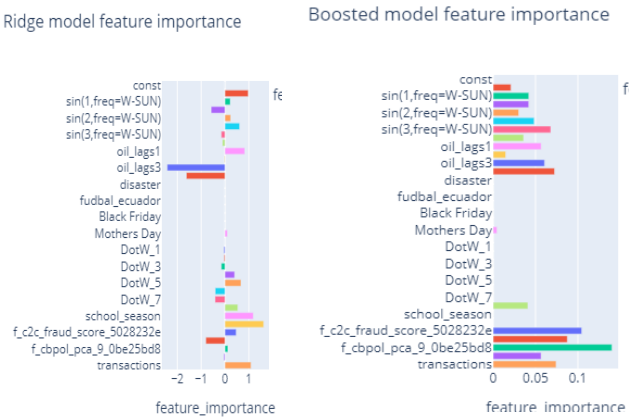
Metrika	Train	Test
RMSLE	/	0.466
RMSE	128.64	/

hibridni model



Hibridni model store 52 product PET SUPPLIES

Sa grafika vidimo da je model savršeno predvideo prodaju proizvoda PET SUPPLIES iz prodavnice 52, što je u praksi malo verovatno. Tako da za neke prodavnice i proizvode ipak možemo da tvrdimo da je model pretrenirao.

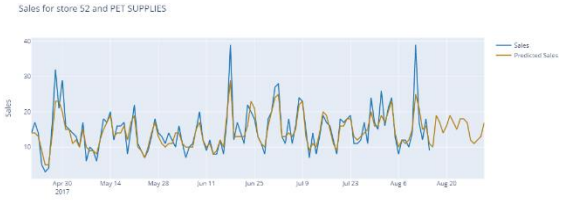


Vaznost hibridnog prediktora za 52 PET SUPPLIES

Sa grafika vidimo da je u Ridge modelu oil_lag3 ima najveću važnost zatim sledi f_holiday_score koja je dobijena iz UpGini biblioteke. Dok je Xgb u rezidualima našao najveću važnost u f_cbpool_pca_9 koja je takođe dobijena iz UpGini biblioteke.

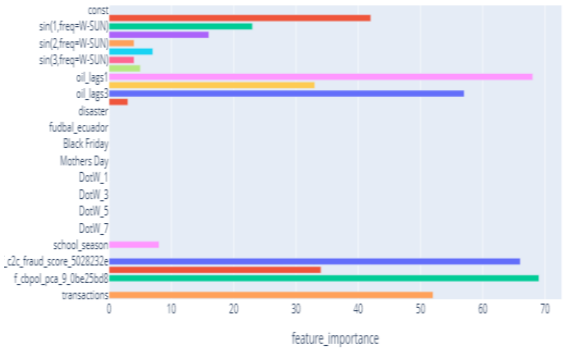
Metrika	Train	Test
RMSE	109.757	/
RMSLE	0.424	0.437

Greške LGBM modela



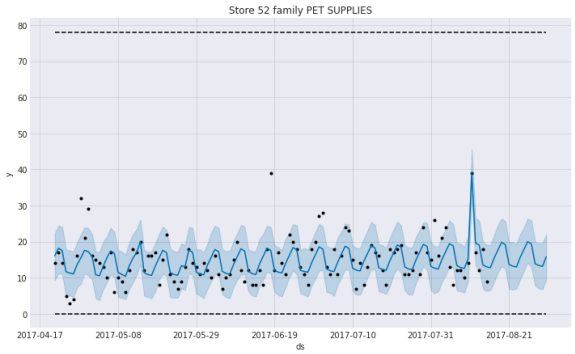
Lgbm model 52 PET SUPPLIES

Sa grafika vidimo da lgbm model nije savršeno predvideo PET SUPPLIES iz prodavnice 52, nije bio overfit. LGBM model bolje predviđa i na test i na train setu. Ako uporedimo iste metrike. Nema ni pretreniranja jer RMSLE ima približnu vrednost i na train i na test setu. Za sada je LGBM naš izbor u odnosu na hibridni model.

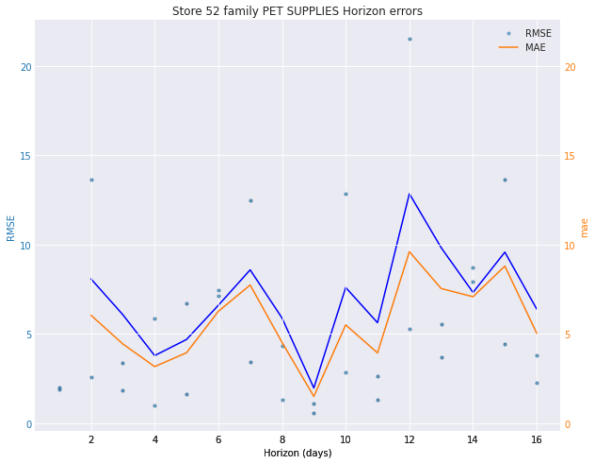


LGBM feature importance 52 PET SUPPLIES

Kao i u xgb u hibridom modelu najveću važnost ima f_cbpool_pca_9, zatim sledi oil lag1.



fbProphet optimized 52 PET SUPPLIES



Metrika	Train	Test
proph_regr	/	0.572
proph_opt_no_regr	/	0.545

fbprophets RMSLE errors

FbProphet je dao nešto lošije rezultate na test setu i ako je probano sa optimizacijom koja je ipak dala nešto bolje rezulate od onog u kome smo ubacili regresore. FbProphet ima dosta opcija koju je bilo moguće modifikovati, ali zbog dugotrajnog učenja i ogromne potrošnje RAM memorije morali smo odabrati samo najvaznije parametre.

```
changepoint_prior_scale': [ 0.01, 0.1, 0.5],
'seasonality_prior_scale': [ 0.1, 1.0, 10.0]
```

U nekim serijama kao što je school&office fbprophet daje nekarakteristične rezultate u oba tipa modela. Naime predviđa na početku ogroman inpuls dok modeli mašinskog učenja ne. Bilo bi zanimljivo videti koji je model više u pravu u tom specifičnom slučaju, ali držaćemo se do toga su ML modeli bolji zbog bolje greške na test setu.

Rezultat predviđanja najboljeg modela LGBM je submitovan na Kaggle takmičenju i trenutno zauzima 76 poziciju!

5 ZAKLJUČAK

Rad na ovom problemu je bio prilično izazovan, pre svega zbog konstantnog feature inžinjiringa. Moglo je još da se eksperimentiše sa dodatnim atributima , na primer da se ubace težine da bliskiji podaci imaju veću važnost. Možda da se ubace eksponencijalni lagovi ili obični iz određenog perioda, dakle širok spektar konstruisanja novih podataka. I ako se čini da se dosta vremena u ovom radu posvetilo baš tome, naprotiv više vremena se provelo u konstruisanju i biranju pravog modela i testiranju novih bliblioteka Dash,UpGini i FbProphet.

Možda bi rezultati bili bolji da sam imao više vremena u testiranju novih atributa.

Ovaj problem tražnje u supermarketima koji je predstavljen u ovom radu, može da bude primenjen univerzalano na bilo koju tražnju. I ako su podaci bili prilično limitirani, na pr ne znamo da li se neki proizvodi uopšte i nalaze u marketu za kojie smo predvideli nule. Kolika je bila promocija? Kada da očekujemo neku novu vrstu robe kao što je bilo sa BOOKS proizvodima itd. Uprkos tome došlo se do zadovoljavajućih rezultata. Mada sam očekvao da će fbprphet dati bolje rezultate, možda da je isprobana optimizacija sa svim regresorima mada bi treniranje tih modela trajalo večno.

Kao što je u uvodu navedeno da je problem bacanja prekomerne hrane supermarketa veliki problem , možda će ovaj rad pomoći da se taj problem mitigira. Bolje predvidi tražnja korisnika za određenim proizvodima i optimizuje skladištenje hrane i drugih potrošnih proizvoda.