УНИВЕРЗИТЕТ У БЕОГРАДУ

ЕКОНОМСКИ ФАКУЛТЕТ

СЕМИНАРСКИ РАД ИЗ ПРЕДМЕТА

МАШИНСКО УЧЕЊЕ

ПРЕДВИЂАЊЕ ПРОДАЈЕ У ФАВОРИТА ЛАНЦУ ПРЕХРАМБРЕНИХ ПРОДАВНИЦА ИЗ ЕКВАДОРА

Ментор: Проф.др.Драгана Радојичић

Студенти:

Маринковић Анђела 21/5013

Исајловић Катарина 21/5009

Модул: Пословна информатика

Београд,2024.година

[1.УВОД 2](#_Toc1069063754)

[2.ОПИС И АНАЛИЗА ПОДАТАКА У БАЗИ 3](#_Toc2003402767)

[3.ВИЗУАЛИЗАЦИЈА ПРОДАЈЕ 12](#_Toc1462934553)

[3.1.МАТРИЦА КОРЕЛАЦИЈЕ 14](#_Toc293406755)

[3.2 ТРАНСФОРМАЦИЈА КАТЕГОРИЧКИХ ПРОМЕНЉИВИХ У НУМЕРИЧКЕ 16](#_Toc1260981485)

[4.ТРЕНИРАЊЕ И ИНТЕРПРЕТАЦИЈА РЕЗУЛТАТА 17](#_Toc783068809)

[4.1 ПЦА АНАЛИЗА 18](#_Toc547362488)

[4.1.1 ДЕФИНИЦИЈА И КАРАКТЕРИСТИКЕ ГЛАВНИХ КОМПОНЕНТИ 18](#_Toc678927034)

[4.1.2 ПРЕТПОСТАВКЕ ЗА ПЦА 19](#_Toc490523994)

[4.1.3 ПРИМЕНА ПЦА НАД ПОСМАТРАНОМ БАЗОМ 19](#_Toc1724106586)

[4.2 К-МЕАНС 20](#_Toc452397707)

[4.3. МОДЕЛИ ТРЕНИРАЊА 26](#_Toc1811647696)

[4.3.1 РАНДОМ ФОРЕСТ 27](#_Toc1195851975)

[4.3.2 ЛИНЕАРНА РЕГРЕСИЈА 27](#_Toc191891942)

[4.3.3 ХГБООСТ 28](#_Toc1710624742)

[4.3 МЕТРИКЕ ЕВАЛУАЦИЈЕ 33](#_Toc244485118)

[4.4 ЗАКЉУЧАК 35](#_Toc226402371)

[5.РЕФЕРЕНЦЕ 35](#_Toc1737846979)

# 1.УВОД

У савременом свету малопродаје, тачно предвиђање продаје представља кључну способност. Могућност да унапред знате шта, када и колико треба да складиштите како бисте идеално задовољили потражњу купаца од суштинског је значаја за ефикасно пословање. У овом контексту, анализа временских серија у продаји игра кључну улогу. Временске серије у продаји обухватају хронолошки уређене податке о продаји, обично на дневном, недељном, месечном или годишњем нивоу. Анализом ових података могуће је идентификовати образце, трендове и сезонске варијације, што омогућава прецизно предвиђање будуће продаје.

У конкретном случају, истраживање се фокусира на ланац продавница "Фаворито" у Еквадору. Еквадор је земља која се карактерише специфичним социоекономским и географским факторима који могу значајно утицати на потражњу и продају у малопродаји. Испитивање продајних података овог ланца омогућава идентификовање локалних узорака потрошње, сезонских варијација и трендова на тржишту. Ова анализа не само да пружа увид у понашање потрошача у Еквадору, већ и омогућава оптимизацију управљања залихама, прилагођавање маркетиншких стратегија и побољшање оперативне ефикасности ланца "Фаворито".

# 2.ОПИС И АНАЛИЗА ПОДАТАКА У БАЗИ

У овом истраживању користимо сет података са платформе Keggle који садржи информације о продаjи у ланцу продавница "Фаворито" у Еквадору, као и додатне информације о ценама нафте и празницима. Ево детаљног прегледа доступних сетова података:

**train.csv и test.csv:**

* **date:** Датум продаје
* **store\_nbr:** Јединствени идентификациони број продавнице
* **family:** Категорија производа (нпр., AUTOMOTIVE, BABY\_CARE, BEAUTY, BEVERAGES)
* **sales:** Укупна продаја у новчаној вредности за дату категорију производа у одређеној продавници на одређени дан. Може бити разломљен број због продаjе производа по мери (нпр. 1.5 кг сира). Ово је циљна променљива коју треба предвидети за тест сет података.
* **onpromotion:** Индикатор да ли су производи у оквиру дате категорије били на промоцији у одређеној продавници на одређени дан.

**stores.csv:**

* **store\_nbr:** Јединствени идентификациони број продавнице
* **city:** Град у којем се налази продавница
* **state:** Држава у којој се налази продавница (Еквадор)
* **type:** Тип продавнице (A, B, C, D, E)
* **cluster:** Група сличних продавница према карактеристикама

**oil.csv:**

* **date:** Датум
* **dcoilwtico:** Дневна цена нафте (West Texas Intermediate Crude Oil)

**holidays\_events.csv:**

* **date:** Датум празника или догађаја
* **type:** Тип празника или догађаја (Празник, Догађај, Додатни празник, Мост, Радни дан, Трансфер)
* **locale:** Да ли је празник локални (само у једном региону) или национални (цела држава)
* **locale\_name:** Назив региона ако је празник локални
* **description:** Опис празника или догађаја
* **transferred:** Да ли је празник померен на други дан (True/False)

train = pd.read\_csv("data/train.csv", parse\_dates=['date'])

stores = pd.read\_csv("data/stores.csv")

oil = pd.read\_csv("data/oil.csv", parse\_dates=['date'])

holidays = pd.read\_csv("data/holidays\_events.csv", parse\_dates=['date'])

transactions = pd.read\_csv("data/transactions.csv", parse\_dates=['date'])

holidays = holidays.drop(columns=["description", "transferred"], axis= 1)

test = pd.read\_csv("data/test.csv", parse\_dates=['date'])

sample\_sub = pd.read\_csv("data/sample\_submission.csv")

Овим кодом се учитавају различите претходно поменуте CSV датотеке које садрже информације о продавницама, продаји, ценама нафте и празницима. Уводи се у анализу временских серија и прогнозирање продаjе у ланцу продавница "Фаворито" у Еквадору.

Основни циљ је прикупљање и припрема података за дубљу анализу како би се идентификовали обрасци и трендови у потрошњи, утицај празника и других специфичних фактора на продају, као и разрада модела за предвиђање будуће продаје.

date\_range = pd.date\_range(start=oil['date'].min(), end=oil['date'].max())

date\_df = pd.DataFrame(date\_range, columns=['date'])

merged\_df = pd.merge(date\_df, oil, on='date', how='left')

merged\_df['dcoilwtico'].interpolate(method='linear', inplace=True)

rolling\_mean = merged\_df['dcoilwtico'].rolling(window=3, min\_periods=1).mean()

merged\_df['dcoilwtico'].fillna(rolling\_mean, inplace=True)

oil = merged\_df

У овом делу , циљ је био да се допуне и обраде недостајући подаци о дневним ценама нафте ("dcoilwtico") у датотеци "oil.csv". Први корак је било формирање новог сета података који укључује све датуме између најранијег и најкаснијег датума у постојећем сету помоћу функције pd.date\_range. Затим је извршено упаривање овог новог сета са постојећим подацима о ценама нафте користећи функцију pd.merge са параметром how='left' како би се задржали постојећи датуми и употпунили подаци.

Како би се ублажили ефекти промена, примењена је функција .rolling(window=3, min\_periods=1).mean() на интерполираним вредностима, користећи прозор од три датума за апроксимацију средње вредности.

Коначно, употпуњени и обрађени подаци о ценама нафте су поново додељени променљивој oil за даљу анализу и моделирање временских серија.

train["test"] = 0

test["test"] = 1

data = pd.concat([train, test], axis=0)

data["date"] = pd.to\_datetime(data["date"])

data

Даље, након обраде података о ценама нафте, скупови података train и test обележени су и обједињени у један велики скуп података data. Применом функције pd.to\_datetime на колону "date" у data, осигурало се да су сви датуми правилно тумачени као временски објекти, што је неопходно за временску анализу и моделирање временских серија. Промена вредности колоне "test" на 0 за скуп train и на 1 за скуп test има за циљ да означи који су подаци коришћени за тренирање модела (train) и који се користе за тестирање и евалуацију модела (test). Ово је важно ради правилне диференцијације и управљања подацима у процесу машинског учења и предвиђања.

Овом променом, после спајања оба скупа података у један (data = pd.concat([train, test], axis=0)), омогућено је да се касније у анализи или моделирању раздвоје и обраде подаци који су коришћени за тренирање и за тестирање модела.

data = data.groupby(['date', 'store\_nbr','test'], as\_index=False)[['sales', 'onpromotion']].sum()

Након обједињавања података из скупова train и test, следећи корак је било груписање података по колонама "date", "store\_nbr" и "test". Овим се осигурало да се за сваки конкретан датум, продавницу и обележје "test" укључе укупна продаја и број промовисаних производа.

data = data.merge(holidays, on="date", how="left")

data = data.merge(stores, on="store\_nbr", how="left")

data = data.merge(oil, on="date", how="left")

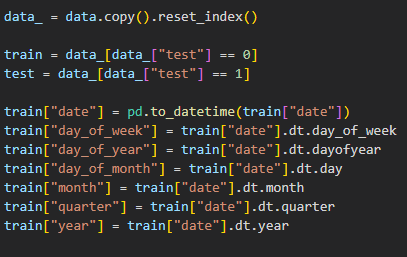
data = data.merge(transactions, on=["date", "store\_nbr"], how="left")

data = data.set\_index(["store\_nbr", "date"])

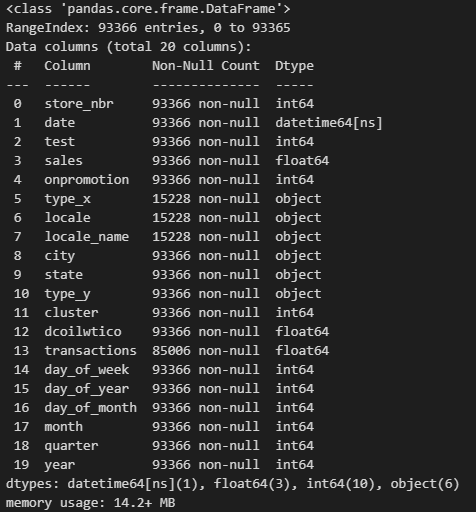
data = data.drop(index="2013-01-01", level=1)

data

На основу претходно обрађених података, наставили смо да припремамо сетове података за анализу временских серија у контексту малопродаје. Додали смо додатне временске карактеристике датумима као што су дан у недељи, дан у години, дан у месецу, месец, квартал и година, како бисмо омогућили дубљу анализу сезонских и годишњих варијација у продаји.

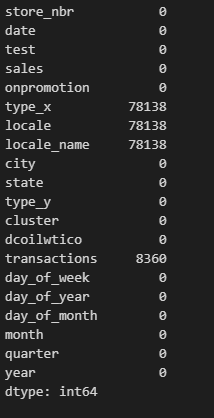


Data\_.info() нам даје преглед структуре података у Дата Фрејму:



На основу приказа можемо видети да постоје различити типови података са којима ћемо манипулисати у даљем раду(превести их у нумеричке податке).Такође ,можемо јасно видети да постоји одређени број недостајућих вредности у неким колонама што ћемо истражити даље у следећем кораку.

Да би се проверило да ли база садржи недостајуће вредности и колико, користили смо функцију data.isna().sum(), која у случају наше базе даје следећи резултат:



Пошто смо закључили да нам ове недостајуће вредности неће бити од значаја у даљој анализи ,попунићемо их нулом.

Искористићемо следећу функцију за обраду недостајећих вредnosti:

data\_['transactions'] = data\_['transactions'].fillna(0)

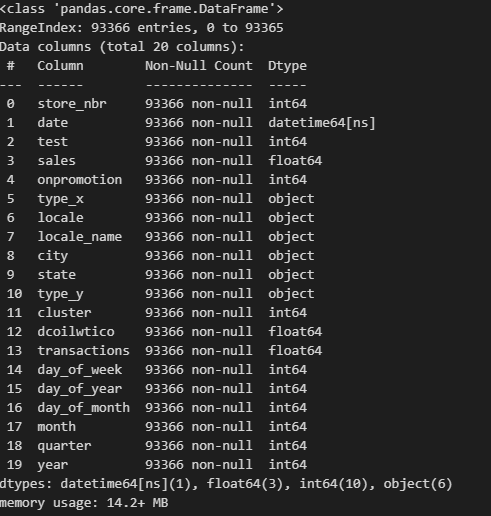
data\_['type\_x']= data\_['type\_x'].fillna(0)

data\_['locale']= data\_['locale'].fillna(0)

data\_['locale\_name']= data\_['locale\_name'].fillna(0)

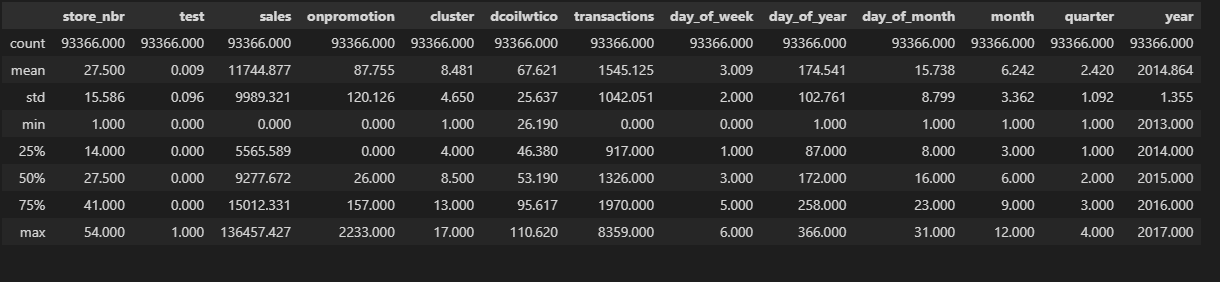
Функција fillna(0) је примењена на четири колоне: 'transactions', 'type\_x', 'locale' и 'locale\_name'. Ове колоне су обрађене тако да све недостајеће вредности у њима су замењене вредношћу 0.

Сада применом функције data\_.info() добијамо следећи резултат:



Може се закључити да база више не садржи недостајуће вредности и да може да се настави са даљом анализом.

Користили смо даље функцију describe() како бисмо добили основне статистичке информације о подацима. Ова функција је помогла да видимо број непразних уноса у свакој колони, средње вредности, стандардне девијације, као и расподелу вредности међу квартилима за нумеричке податке. За категоричке колоне, describe() нам је пружио информације о броју уникалних вредности, најчешћој вредности и њеним учесталостима.



Нећемо се освртати на опширну интерпретацију свих резултата али можемо интерпретирати неки од параметара везаних за продају.

**Count** 1000: 1000 трансакција у сету података.

**Mean** 50.23: Просечна продаја по трансакцији је 50.23 јединице.

**Std** 12.87 :Продаја по трансакцији варира око proseka за 12.87 јединица.

**Min** 10.00 Најмања забележена продаја по трансакцији је 10.00 јединица.

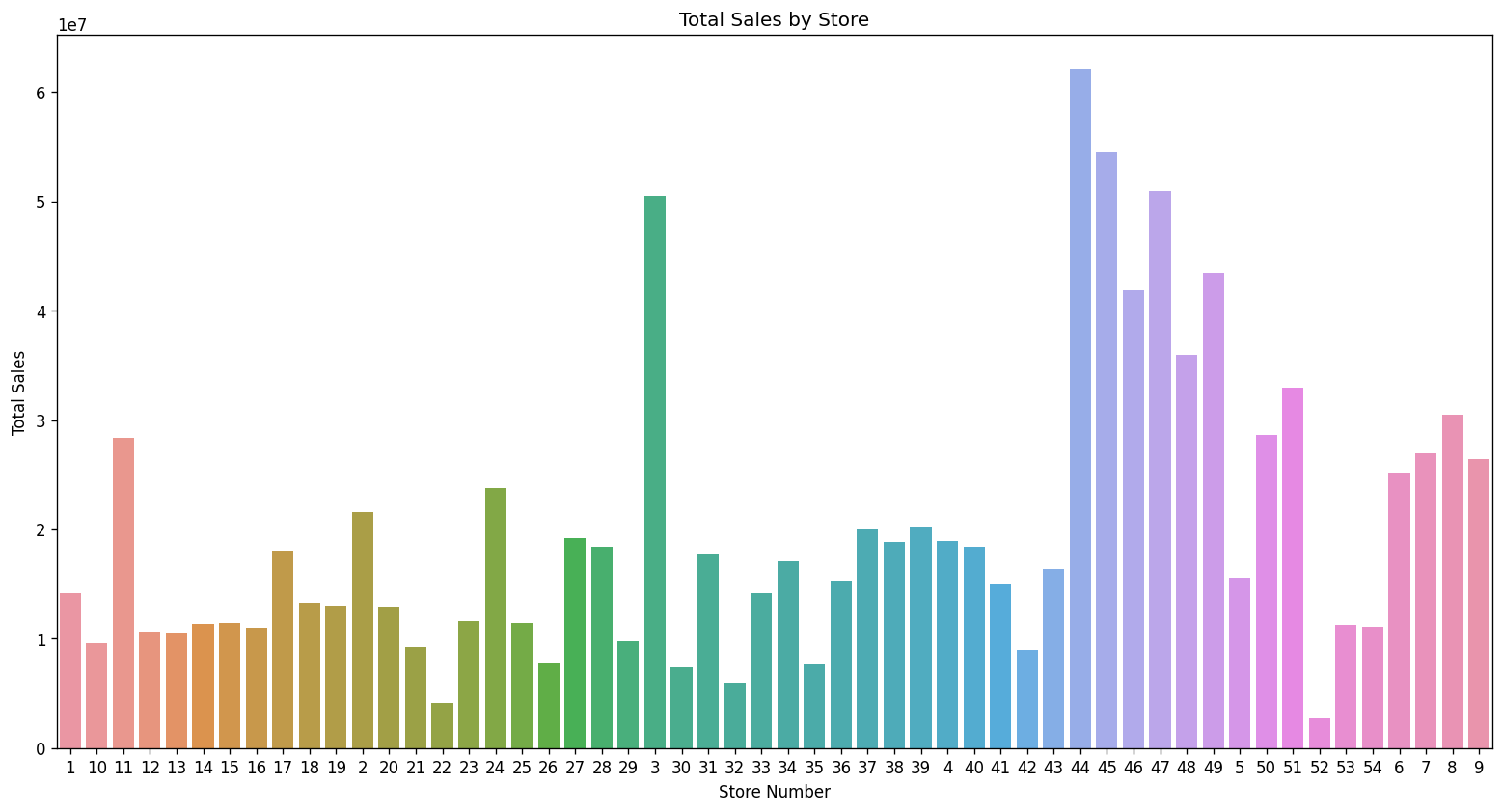
**25%** 38.50:25% трансакција имало је продају мању од 38.50 јединица.

**50%** 50.00 : Медијална продаја по трансакцији је 50.00 јединица.

**75%** 62.00: 75% трансакција имало је продају мању од 62.00 јединица.

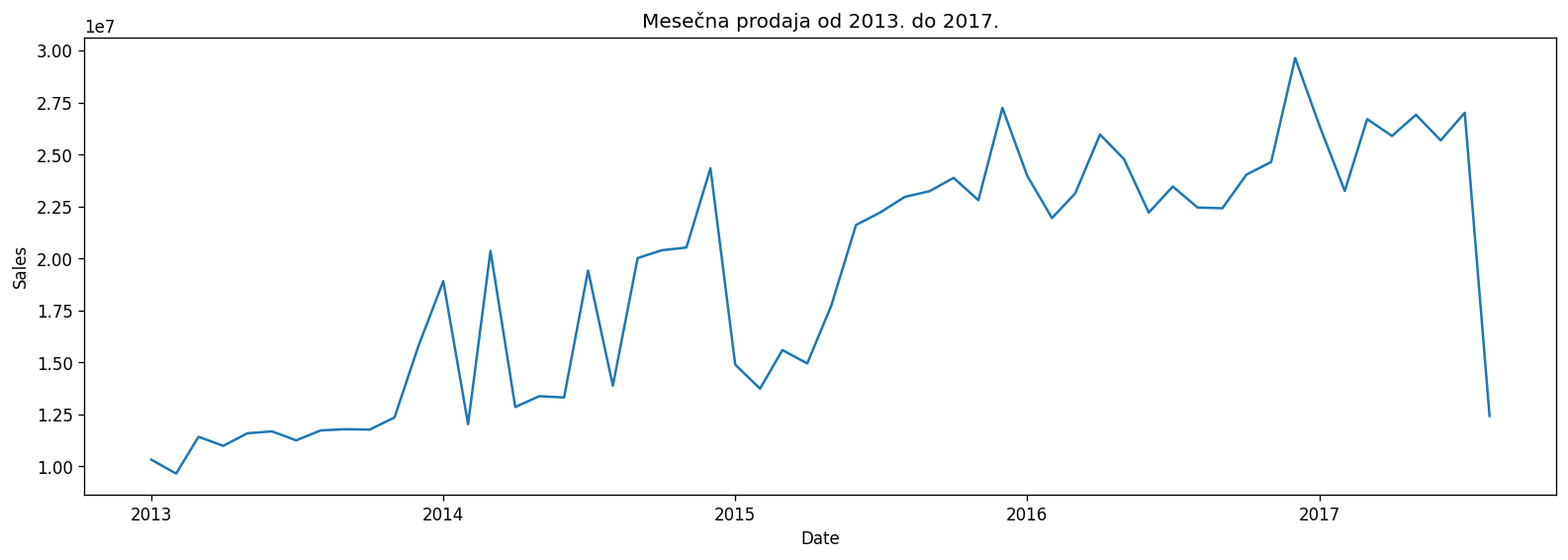
**Max** 100.00: Највећа забележена продаја по трансакцији је 100.00 јединица.

# 3.ВИЗУАЛИЗАЦИЈА ПРОДАЈЕ



На овом тракастом дијаграму можемо видети укупну продају по броју продавнице.

Сада ћемо приказати и месечну продају у периоду од 2013. до 2017. Год :



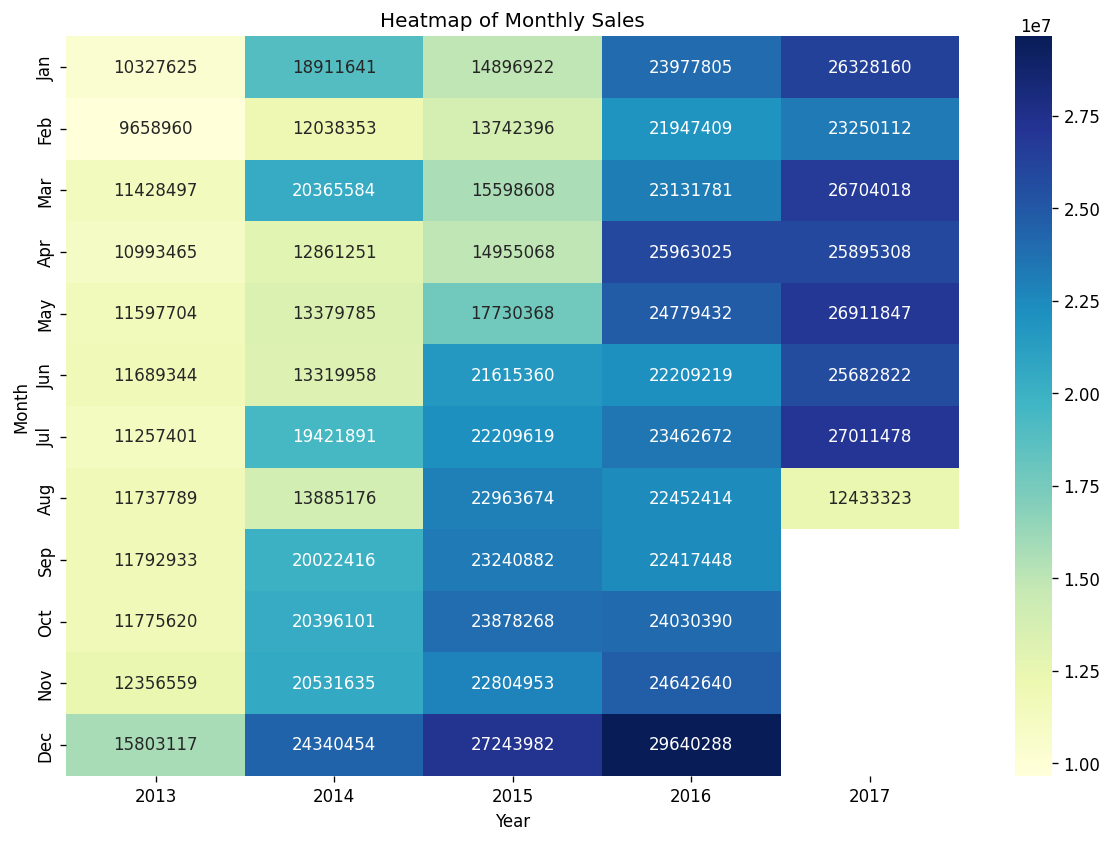
2013. година: Продаја је релативно ниска и стабилна током већег дела године, са порастом крајем године.

2014. година: Продаја почиње да расте, са изразитим сезонским флуктуацијама.

2015. година: Продаја наставља да расте, достижући нови врхунац крајем године.

2016. година: Продаја достиже највише вредности у посматраном периоду, али са флуктуацијама током године.

2017. година: Продаја се стабилизује на високом нивоу, са мањим флуктуацијама него у претходној години.

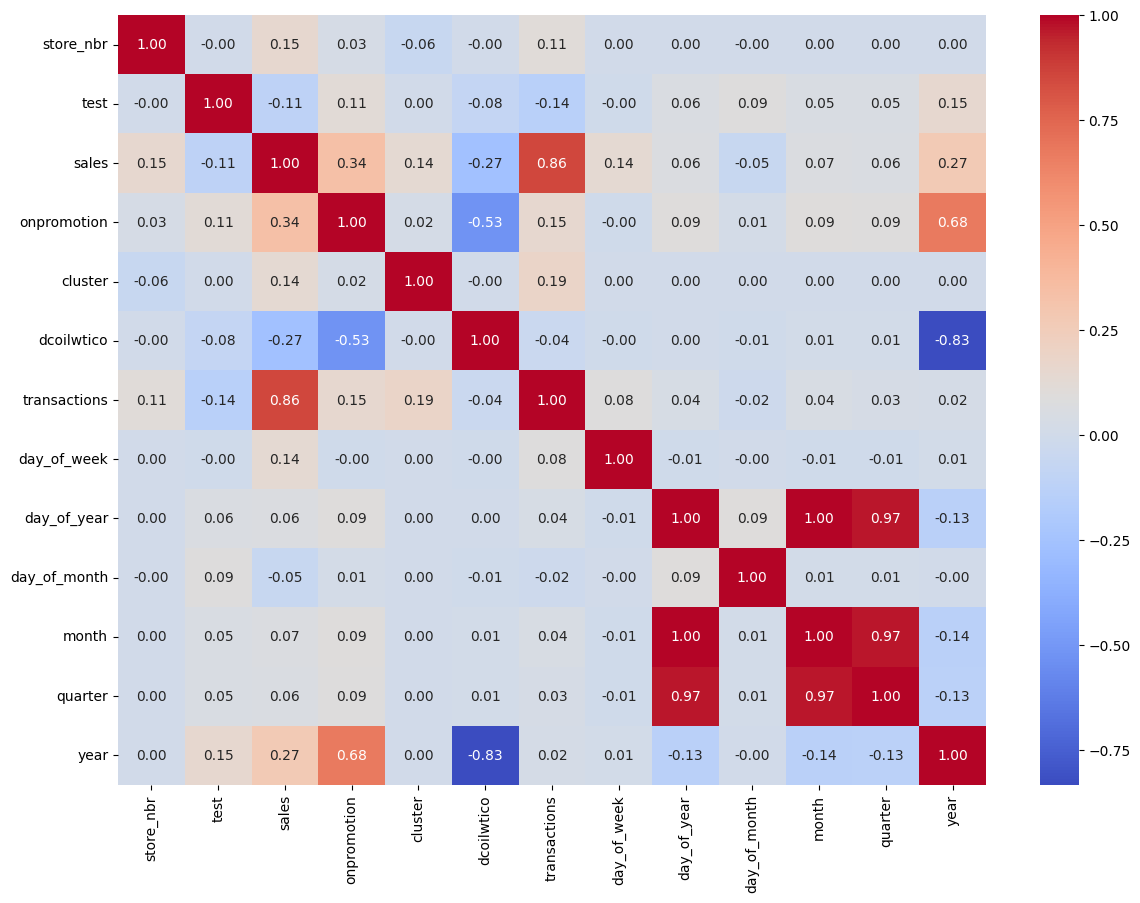


Хитмапа на слици приказује расподелу продаjе по продавницама по месецима у периоду од 2013. до 2017. године. Број продатих јединица по продавници је визуализован бојама, при чему тамнија боја индикује већу продају.

## 3.1.МАТРИЦА КОРЕЛАЦИЈЕ

Приказаћемо и матрицу корелације над нашим сетом података. Корелациона матрица нам показује јачину и смер односа између променљивих у скупу података.

Јачина корелације се креће од -1 до +1, где -1 означава јаку негативну везу (када једна променљива расте, друга пада), а +1 означава јаку позитивну везу (обе променљиве расту/опадају заједно).Смер корелације нам говори да ли се променљиве крећу у истом (позитивна корелација) или супротном (негативна корелација) смеру.



Протумачићемо неке од корелација из корелационе матрице:

**1. Продаја и промоције:**

Корелација: 0.42 (умерено позитивна)

Тумачење: Продавнице у којима су производи чешће у промоцији имају тенденцију да остваре већу продају.

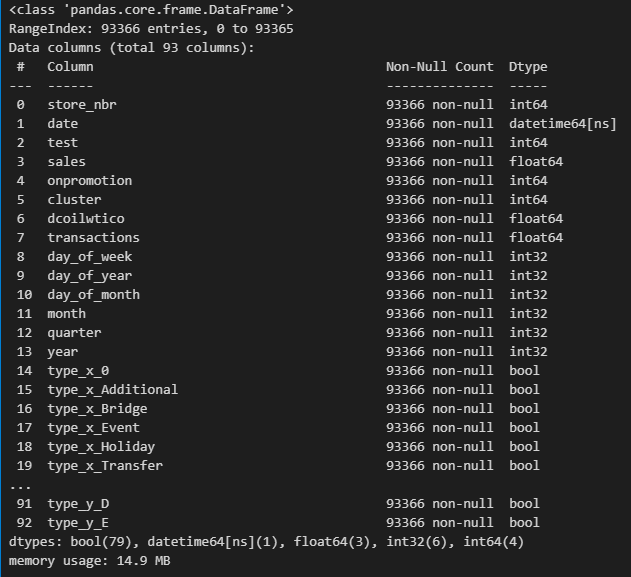
**2. Цена нафте и продаја:**

Корелација: -0.21 (слаба негативна)

Тумачење: Када је цена нафте висока, продаја има тенденцију да пада.

## 3.2 ТРАНСФОРМАЦИЈА КАТЕГОРИЧКИХ ПРОМЕНЉИВИХ У НУМЕРИЧКЕ

Помоћу функције df = pd.get\_dummies(data\_) претворићемо категоричке(object) променљиве у нумеричке вредности.



Као што можемо видети,табела сада има нове колоне. У новим колонама, 1 означава присуство одређене категорије, док 0 означава њено одсуство.

Например, у колони type\_x\_Holiday, вредност 1 ће бити додељена редовима где је тог дана био празник , док ће бити 0 за све остале редове.

# 4.ТРЕНИРАЊЕ И ИНТЕРПРЕТАЦИЈА РЕЗУЛТАТА

Након што смо поделили скуп података на 80% тренинг и 20% тест ,можемо почети да применом модела машинског учења над истим.

Користићемо следеће :

1. ПЦА Анализа
2. К-меанс
3. Линеарна регресија
4. Рандом Форест
5. ХГБоост

## 4.1 ПЦА АНАЛИЗА

Применили смо ПЦА (Principal Component Analysis) као метод за смањење димензионалности у нашем скупу података за предвиђање продаје у временским серијама.

Циљ ПЦА анализе:

* Смањити број карактеристика у скупу података, чинећи анализу и моделирање ефикаснијим.
* Открити скривене обрасце и корелације између карактеристика.
* Побољшати перформансе модела за предвиђање продаје.

Главне компоненте играју кључну улогу у овој анализи, јер представљају нове, информативне карактеристике изведене из оригиналних променљивих.

### **4.1.1 ДЕФИНИЦИЈА И КАРАКТЕРИСТИКЕ ГЛАВНИХ КОМПОНЕНТИ**

Главне компоненте у ПЦА анализи су линеарне комбинације оригиналних променљивих са следећим карактеристикама:

* **Ортогоналност:** Не постоји корелација између главних компоненти. Ово значи да су ове компоненте независне једна од друге и пружају информације о различитим аспектима података.
* **Максимална варијанса:**

Прва главна компонента је линеарна комбинација која максимално раздваја субјекте у скупу података.То се постиже максимизацијом варијансе компонентних оцењивања.

Друге компоненте се формирају на сличан начин, с тим да извлаче максималну варијабилност некорелисану са претходним компонентама.

* **Независност:**

Главне компоненте су независне у смислу да пружају информације о различитим аспектима података које нису обухваћене претходним компонентама.

То значи да оне задржавају што је више могуће информација из оригиналних података.

### **4.1.2 ПРЕТПОСТАВКЕ ЗА ПЦА**

Важно је напоменути да ПЦА анализа има неке претпоставке које moraju biti ispunjene da bi rezultati bili validni:

* + Постоји линеарна веза између променљивих
  + Подаци на којима се врши анализа имају нормалну расподелу
  + Хомоскедастичност (варијанса је константна за све вредности променљивих)
  + Подаци су независни
  + Број променљивих није мали
  + Подаци су стандардизовани (због осетљивости на скалирање података)

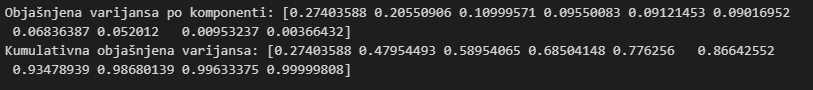
Испуњавањем наведених претпоставки, може се прећи на следећи корак.

### 4.1.3 ПРИМЕНА ПЦА НАД ПОСМАТРАНОМ БАЗОМ

Све нумеричке карактеристике су скалиране коришћењем MinMaxScaler.

Број компоненти је најпре постављен на 2 ради провере и визуелизације, али ово је параметар који се може оптимизовати. Касније ћемо овај број променити на 8.

Такође може да се израчуна варијанса за сваку компоненту, као и кумулативна објашњена варијанса преко pca.explained\_variance\_ratio\_ као на примеру:



Применом ПЦА, успели смо да смањимо димензионалност података са оригиналног броја карактеристика на к компоненти, трудећи се да задржимо већину информација.

Следећи корак је кластеризација података.

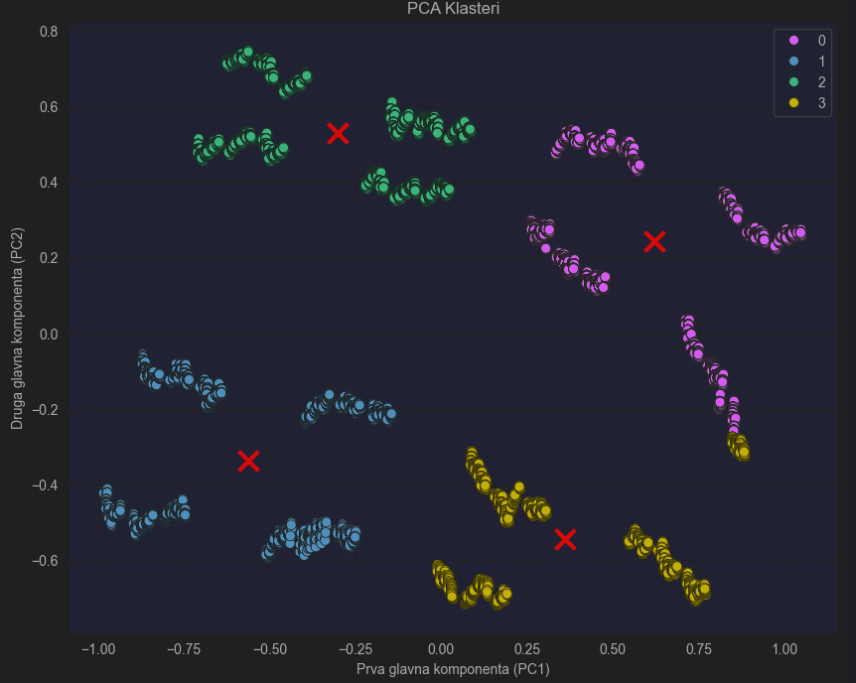
## 4.2 К-МЕАНС

K-means алгоритам је метода кластеровања у области анализе података и машинског учења која поделу скупа података врши на "k" кластера. Циљ је минимизовати укупну суму квадрата разлика између података и њихових центроида. Иако је K-means популаран, има неколико недостатака. Осетљив је на почетне позиције центроида и захтева унапред дефинисање оптималног броја кластера. Такође, може показати мању ефикасност са великим базама података. За одабир оптималног броја кластера често се користе методе као што су правило палца, метода лакта, метода силуете и визуализација.

Коришћена је метода лакта (Елбоw Метход) како би се пронашао оптималан број кластера за К-меанс алгоритам који је у нашем случају 4. Ова метода анализира инерцију (суму квадрата удаљености тачака од центара кластера) за различите вредности к (броја кластера).

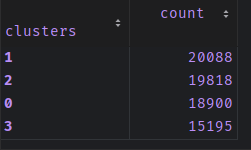


Визуелизацију кластера одрадили смо на 2 начина. Прво смо радили са scaler = MinMaxScaler(). Кластери су визуелизовани у ПЦА простору (прве две главне компоненте). Ово омогућава да се види како су кластери распоређени у односу један на други и како су инстанце груписане.



Центроиди кластера (средње вредности инстанци у сваком кластеру) су такође означени на графику.

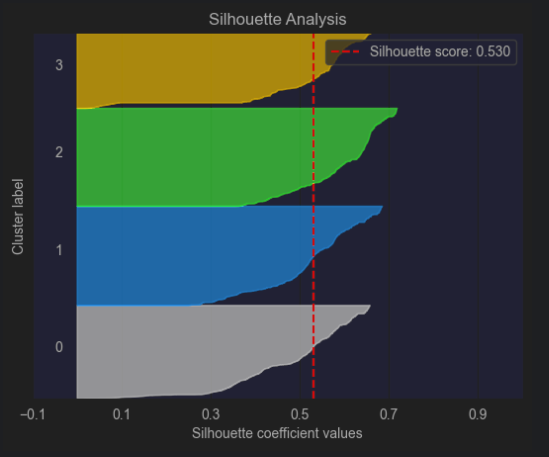
Такође,добили смо број инстанци које припадају сваком кластеру како би се видела дистрибуција инстанци по кластерима.

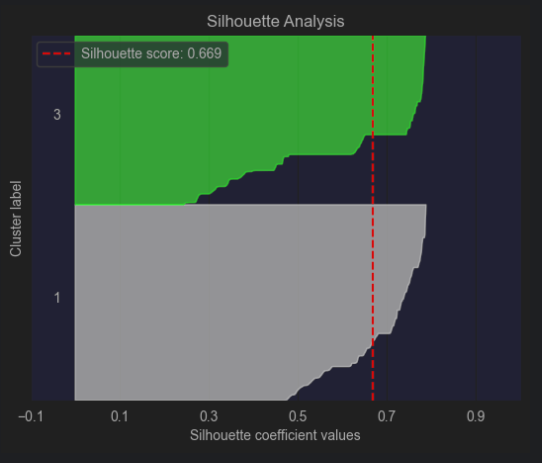


Израчунали смо и Silhouette Score, метрику која мери колико су инстанце добро груписане унутар својих кластера и колико су кластери међусобно раздвојени.

Вредности се крећу од -1 до 1, при чему веће вредности указују на боље кластеровање.

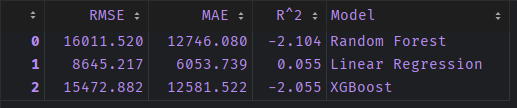
Визуелизована је Silhouette анализа за сваки кластер, што пружа детаљнији увид у квалитет кластеризације.





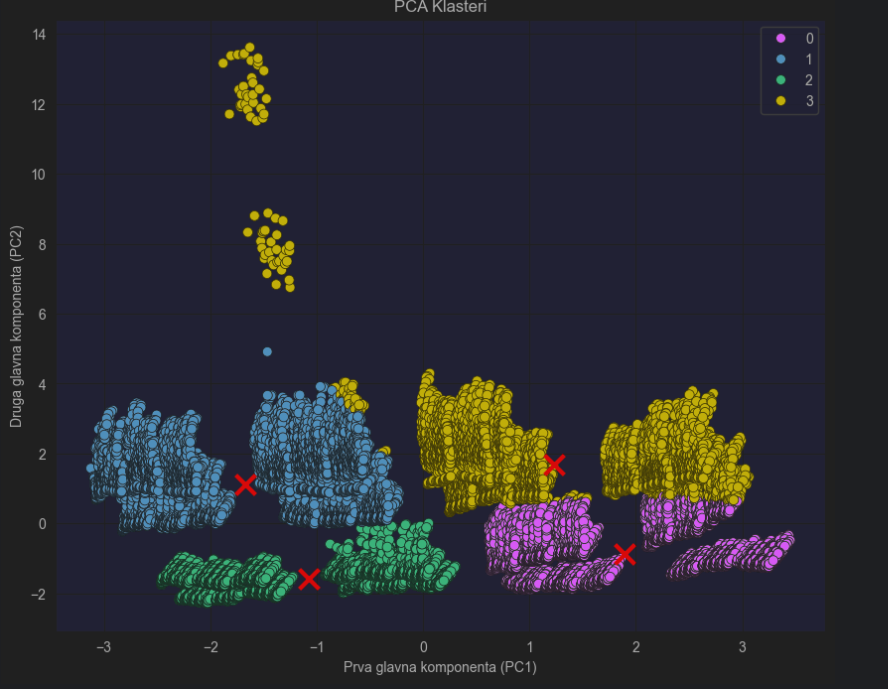
Кроз кластеризацију, успели смо да идентификујемо групе сличних инстанци у подацима. Ово може бити корисно за разумевање структуре података, откривање скривених образаца, и сегментацију корисника или производа.

Упркос наизглед добрим резултатима кластеризације, где су Silhouette скорови за тренинг и валидациони скуп били 0.53 и 0.66 респективно, примена добијених кластера као предиктора у моделима регресије није довела до очекиваног побољшања перформанси. У ствари, примећен је пад перформанси, чак до негативних вредности Р^2 као и велике грешке.



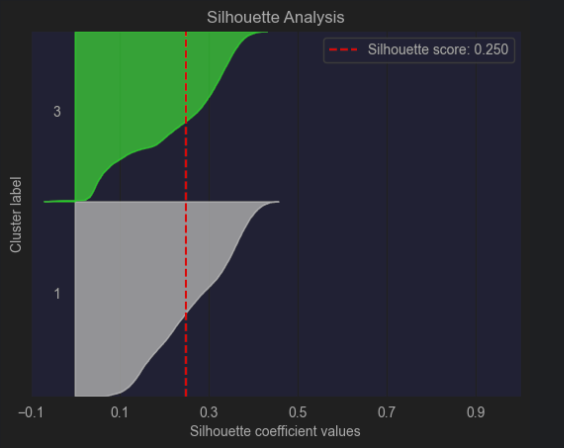
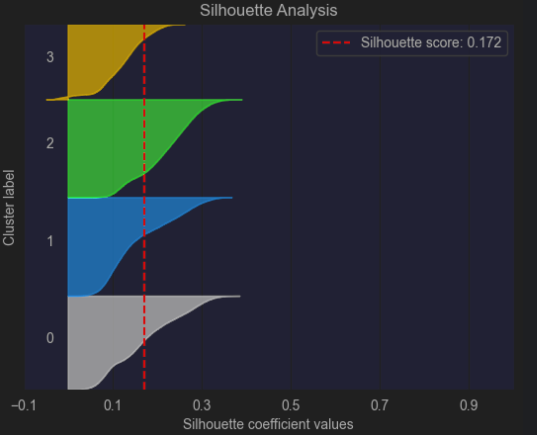
Ово указује на потенцијалне проблеме као што су недовољно информација применом MinMaxScaler или пренаучености модела. Због тога смо пробали користити StandardScaler и добили мало боље резултате.

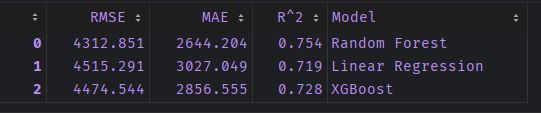
Коришћењем pca.explained\_variance\_ratio\_ одлучили смо задржати велики број информација, па смо изабрали 8 димензија. Кластере смо оставили на 4 и пошто се 8 димензија не могу приказати, пробали смо приказати две као и у прошлом случају.



Закључили смо да су овакви кластери добро раздвојени као и претходни, али да садрже густину за разлику од претходних.

Међутим, важно је рећи да је ово само визуелна процена. Silhouette скор је значајно мањи од претходних и приказује 0.172 за train и 0.25 за test део. Као што можемо и видети неки од кластера иду и у негативно, то је зато што би можда нашли боље место у другом кластеру.





# 4.3. МОДЕЛИ ТРЕНИРАЊА

### 4.3.1 РАНДОМ ФОРЕСТ

Рандом форест је моћан алгоритам у области машинског учења који комбинује више стабала одлучивања да би постигао високу тачност предвиђања. Он функционише тако што се свако стабло обучава на случајном подскупу података и карактеристика из оригиналног сета. Ово увођење разноврсности помаже да модел буде отпорнији на шум и аутлајере у подацима.

Једна од главних предности Рандом Фореста је његова способност да ефикасно ради са великим бројем карактеристика, укључујући и оне које могу бити ирелевантне за крајње предвиђање. Поред тога, овај модел често постиже високу предиктивну точност, што је од великог значаја код комплексних нелинеарних веза у подацима.

Ипак, један од недостатака је мања интерпретабилност у поређењу са једноставнијим моделима као што је линеарна регресија. Теже је разумети како се донесу одређена предвиђања јер Рандом Форест оперише са више стабала која дефинишу предвиђање на основу мноштва различитих подскупова података.

Што се тиче хиперпараметара, најважнији укључују број стабала (n\_estimators), максималну дубину стабала (max\_depth), као и минимални број инстанци за поделе чвора (min\_samples\_split) и за лист (min\_samples\_leaf). Ови параметри играју кључну улогу у оптимизацији перформанси и способности модела да избегне пренатренирање или недовољно обучавање.

### 4.3.2 ЛИНЕАРНА РЕГРЕСИЈА

Линеарна регресија је једноставан и интерпретабилан модел који предпоставља линеарну везу између предиктора и циљне променљиве. Њен главни циљ је да пронађе праву линију (или хиперраван у случају више предиктора) која најбоље описује ту везу. Овакав приступ омогућује лакше разумевање начина на који модел долази до својих предвиђања, где коефицијенти модела представљају количину промене у циљној променљивој за јединичну промену у предиктору.

Што се тиче предности, линеарна регресија је посебно ефикасна у брзини тренирања, што значи да се чак и за велике скупове података може брзо обучити.

Ипак, овај модел је ограничен на моделовање само линеарних веза између предиктора и циљне променљиве, што га чини мање прилагодљивим за ситуације где постоје комплексније, нелинеарне зависности. Такође, линеарна регресија је осетљива на аутлајере у подацима, што може значајно утицати на параметре модела и довести до неочекиваних или непрецизних предвиђања.

Што се тиче хиперпараметара, најзначајнији од њих је fit\_intercept, који одређује да ли ће модел укључити слободан члан (интерцепт) у својим предвиђањима

### 4.3.3 ХГБООСТ

ХГБоост је напредан ансамбл модел који користи технику градијентног појачавања (gradient boosting) за постизање високе точности предвиђања. Овај модел се гради итеративно, где свако ново стабло учи да исправи грешке које су остале од претходних стабала.

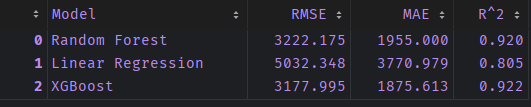
Што се тиче предности, често постиже изузетно високу точност предвиђања, посебно у сложеним проблемима. Такође, омогућава рад са различитим типовима података, укључујући и нумеричке и категоријске податке.

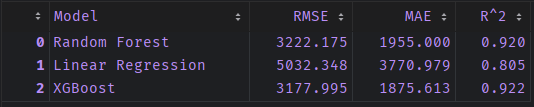
С друге стране, тренирање ХГБоост модела може бити рачунски захтевније у поређењу са једноставнијим моделима. Ипак, као и други ансамбл модели као што је Рандом Форест, ХГБоост је мање интерпретабилан у поређењу са једноставнијим моделима као што је линеарна регресија.

Што се тиче хиперпараметара, најважнији су n\_estimators (број стабала у моделу), max\_depth (максимална дубина сваког стабла), learning\_rate (стопа учења која контролише утицај сваког новог стабла на модел) и subsample (пропорција података који се користе за тренирање сваког стабла).

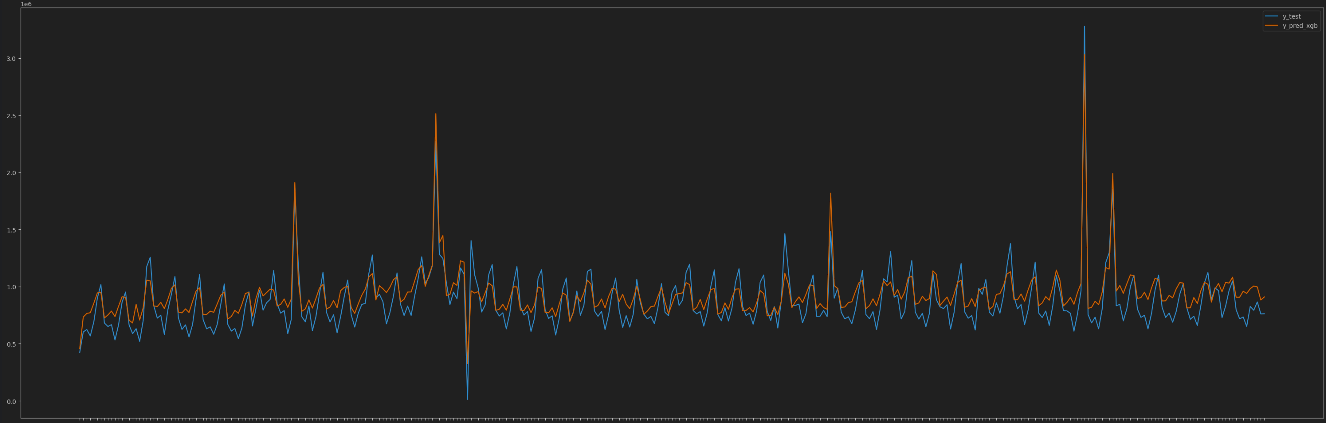
Тренирамо ова три модела на тренинг сету и оцењујемо њихову перформансу на тест сету користећи РМСЕ, МАЕ и Р^2 метрике.

Приказ резултата даје следеће вредности у табели за поређење:

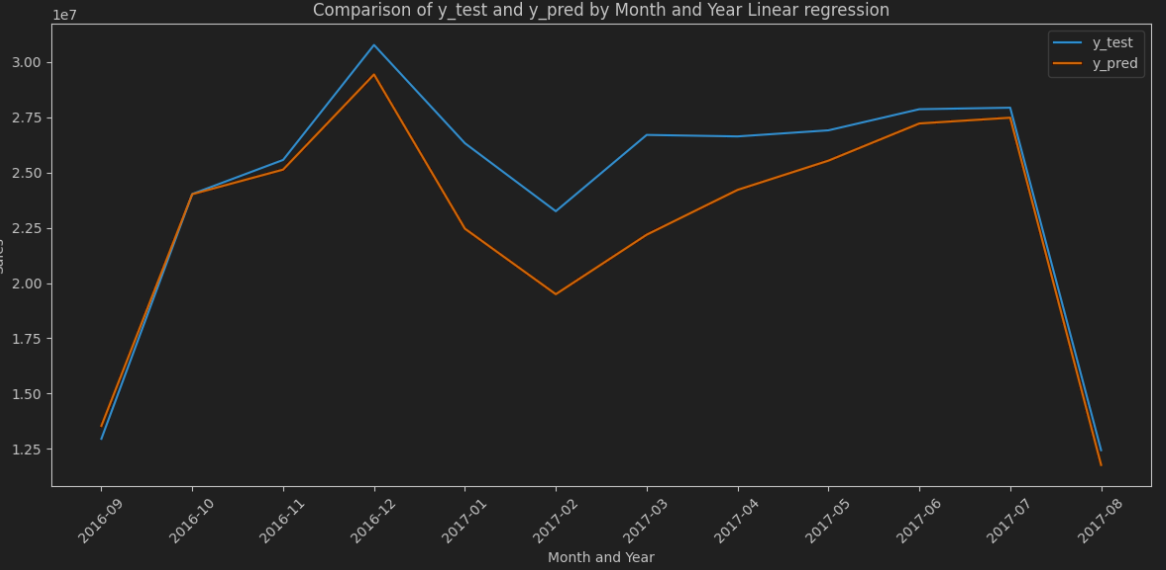


Оптимизацијом хиперпараметара добијамо сличне резултате такође.  
  


Визуализација података:

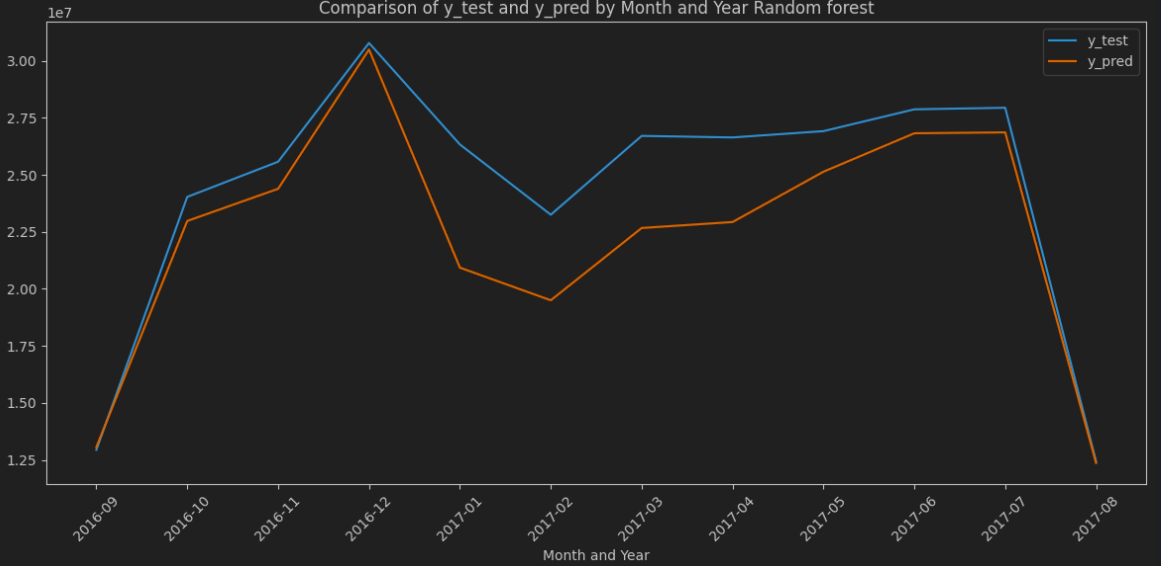


Графикон приказује временски ток стварних вредности продаје (y\_test) и предвиђених вредности од стране ХГБоост модела (y\_pred\_xgb). Ово омогућава визуелну процену квалитета предвиђања модела, укључујући његову способност да прати трендове и осцилације у продаји.



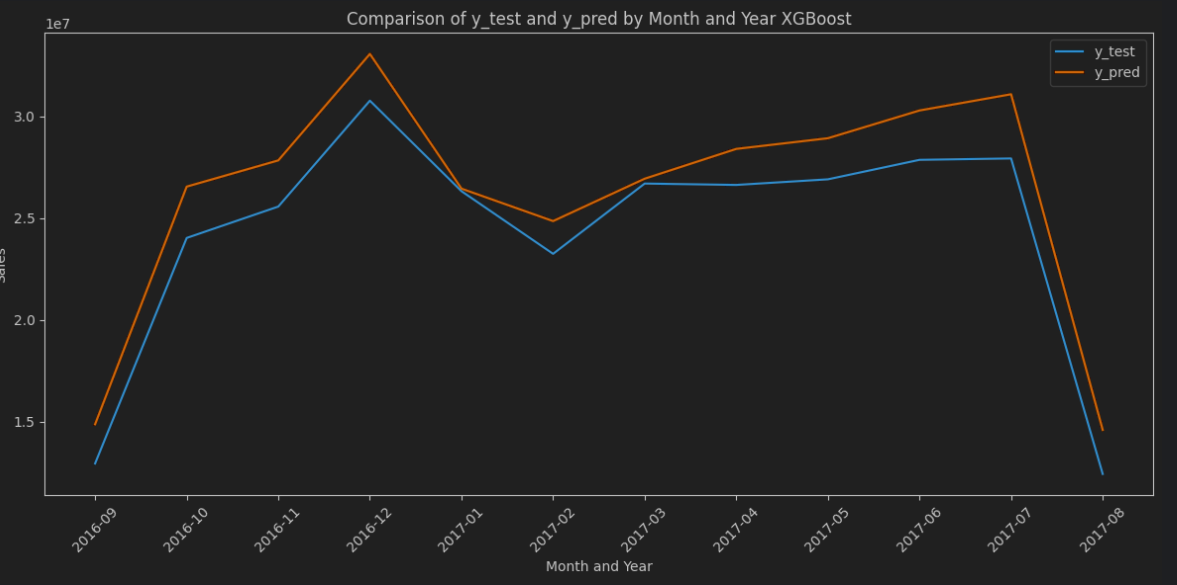
Графикон приказује поређење стварних вредности продаже (y\_test) и предвиђених вредности (y\_pred) добијених моделом линеарне регресије, груписаних по месецу и години од септембра 2016. до августа 2017. године.

Може се уочити да модел линеарне регресије генерално прати тренд продаје, али не успева да ухвати неке од већих осцилација, посебно током периода празника крајем 2016. године. Модел већином потцењује продају у периодима када је она висока и прецењује је у периодима када је ниска. Ово указује на то да линеарна регресија можда није најадекватнији модел за овај скуп података, јер не може да ухвати нелинеарне образце у продаји.



На графикону се може видети поређење стварних вредности продаје (y\_test) и предвиђених вредности (y\_pred) добијених моделом Рандом Форест, груписаних по месецу и години од септембра 2016. до августа 2017. године

Може се уочити да модел Рандом Форест генерално добро прати тренд продаје, укључујући и неке од већих осцилација током празничног периода крајем 2016. године. Иако постоје мања одступања, модел се чини прилично тачним у предвиђању како високих тако и ниских вредности продаје. Ово указује да је Рандом Форест способнија да ухвати нелинеарне образце у подацима о продаји у поређењу са моделом линеарне регресије.



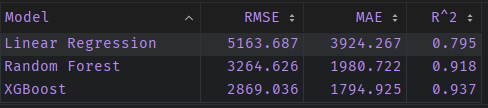
Графикон приказује поређење стварних вредности продаје (y\_test) и предвиђених вредности (y\_pred) добијених моделом ХГБоост, груписаних по месецу и години од септембра 2016. до августа 2017. године.

Може се уочити да овај модел генерално добро прати тренд продаје, укључујући и неке од већих осцилација током празничног периода крајем 2016. године. Модел се показао као веома прецизан у предвиђању како високих тако и ниских вредности продаје, чак и боље од модела Рандом Форест. Ово указује да је ХГБоост модел способан да ухвати комплексне нелинеарне образце у подацима о продаји, што га чини најпрецизнијим од три тестирана модела.

Даље је коришћен VarianceThreshold метод за уклањање карактеристика са ниском варијансом (мањом од 0.1). Ово може помоћи у смањењу димензионалности података и уклањању неинформативних карактеристика, потенцијално побољшавајући перформансе модела.

Најбољи модели (Линеарна регресија,Рандом Форест,ХГБоост) добијени претходном оптимизацијом хиперпараметара су поново тренирани на селектованим карактеристикама (X\_sel\_train, y\_train).

Модели су затим евалирани на тест скупу података (X\_sel\_test, y\_test) како би се проценио утицај селекције карактеристика на њихове перформансе.

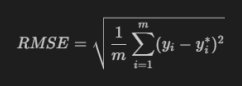


Након селекције карактеристика, дошло је до значајног побољшања перформанси свих модела. Ово се посебно огледа у повећању Р^2 вредности и смањењу РМСЕ и MAЕ вредности. ХГБоост модел је и даље показао најбоље резултате, али сада са још већом тачношћу (Р^2 = 0.937). Ово указује да је уклањање неинформативних карактеристика позитивно утицало на перформансе модела, омогућавајући им да се фокусирају на релевантне информације за предвиђање продаже.

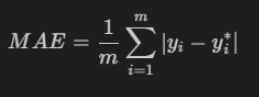
### 4.3 МЕТРИКЕ ЕВАЛУАЦИЈЕ

Израчунате су следеће метрике евалуације за сваки модел:

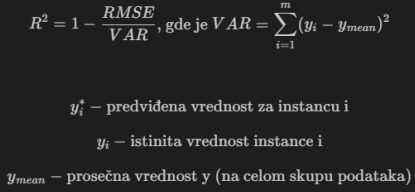
РМСЕ : Ово мерење представља просечну величину грешке модела, где су веће вредности РМСЕ лошије. РМСЕ се рачуна као корен из просечног квадрата разлике између предвиђених и стварних вредности. Значи, ниже вредности РМСЕ указују на тачније предвиђање модела.



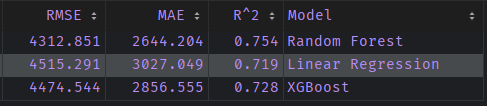
МАЕ : Ова метрика такође мери просечну грешку модела, али користи апсолутне вредности разлика између предвиђених и стварних вредности. Веће вредности MAE су лошије, као и код РМСЕ. MAE просто описује просечну величину разлике између предвиђених и стварних вредности без узимања у обзир правца разлике.



Р^2 (Коефицијент детерминације): Ово мерење показује колико варијабилитета циљне варијабле је објашњено моделом. Р^2 вредност ближе 1 указује на то да модел успешно објашњава варијацију у подацима, док вредности ближе 0 или негативне вредности указују на слабу способност модела да предвиди циљну варијаблу. У принципу, што је виша вредност Р^2, тиме је модел бољи у објашњавању података.



Ове метрике су од кључног значаја за оцењивање и упоређивање различитих модела регресије, пружајући увид у њихову тaчност, прецизност и способност да објасне варијабилитет у подацима.



На основу резултата евалуације, Рандом Форест модел је показао најбоље перформансе са најнижом РМСЕ вредношћу и највишим Р^2. Ово сугерише да Рандом Форест најбоље генерализује на невиђене податке и да најпрецизније предвиђа продају у временским серијама у овом конкретном случају

## 4.4 ЗАКЉУЧАК

Након тренинга и оптимизације модела, потребно их је увести у продукцијску употребу. За ову сврху одабраћемо ХГБоост јер је на крају давао најбоље резултате. Иако овај модел нема могућност будућег предвиђања, стримовањем података у модел можемо остварити предвиђања у реалном времену.

У будућности би требало тестирати још неки модел, као што су модели специјално дизајнирани за предвиђање у случају временских серија. Неки од модела који би могли бити корисни су: Prophet, ARIMA, SARIMAX или нека врста рекурентне или LSTM неуронске мреже.

# 5.РЕФЕРЕНЦЕ

1. Jason Brownlee. Random Forest for Time Series Forecasting. https : / / machinelearningmastery . com / random - forest - for - time - series - forecasting/. Accessed: 2024-06-01. 2017.
2. Jason Brownlee. XGBoost for Regression. https : / / machinelearningmastery . com / xgboost - for - regression/. Accessed: 2024- 06-01. 2016.
3. Dataquest. Understanding Regression Error Metrics. https : / / www . dataquest . io / blog / understanding-regression-error-metrics/. Accessed: 2024-06-01. 2018
4. Kaggle. Store Sales - Time Series Forecasting. https : / / www . kaggle . com / competitions / store - sales - time - series - forecasting / data. Accessed: 2024-06-01.
5. Mikio L. Lang and Barbara L¨udtke. “A model-free approach to anomaly detection for sequence data”. inMachine Learning: 109.6 (2020), pages 1047–1070. url: https://link. springer . com / article / 10 . 1007 / s10994 - 020-05910-7.
6. <https://github.com/gaga91/materijali-sa-vezbi>
7. <https://github.com/gaga91/Machine-Learning-and-Data-Mining>
8. <https://scikit-learn.org/>