Проект по Вовед во науката за податоци

Сања Крстеска 211046

Collect sales data from 3 different sources and make a sales forecasting model

GitHub link:

https://github.com/sanjakrsteska/VNP_project

Video link:

https://www.youtube.com/watch?v=FrCSCseqxQc&ab_channel=SanjaKrsteska

Содржина

Тема: Собирање податоци за продажба од 3 различни извори и градење модел за предвидување на продажбата

Апстракт

Овој труд претставува сеопфатен пристап за развој на модел за прогнозирање на продажбата преку интегрирање на податоци од три различни извори. Методологијата опфаќа собирање податоци, чистење, истражувачка анализа, инженерство на карактеристики, избор на модели, обука и евалуација. Избраниот модел е обучен на податоци за продажба и се оценува со користење на стандардни метрики. Моделот ја зема во предвид корелацијата помеѓу рекламите на телевизија и радио и продажбата. Истиот врши предвидување на идните продажби во зависност од рекламите на ТВ и радио.

Вовед

Трошоците за рекламирање се основна компонента на маркетинг стратегијата, насочена кон зголемување на свесноста за брендот, генерирање на потенцијални клиенти и на крајот поттикнување на продажбата. Разбирањето на врската помеѓу рекламните инвестиции и продажните резултати е од клучно значење за оптимизирање на буџетите за маркетинг и максимизирање на повратот на инвестицијата. Додека претходните студии ја испитуваа ефективноста на рекламирањето низ различни медиумски канали, овој труд конкретно се фокусира на влијанието на телевизиските (ТВ) и радио рекламите врз продажбата.

Прогнозирањето на продажбата игра клучна улога во деловното планирање и процесите на донесување одлуки. Точните предвидувања за идните продажби им овозможуваат на организациите да го оптимизираат управувањето со залихите, распределбата на ресурсите и стратешкото планирање. Сепак, точното предвидување на продажбата може да биде предизвик поради динамичната природа на пазарните услови, сезонските варијации и сложеноста на основните фактори кои влијаат на однесувањето на потрошувачите.

Методологија

2.1 Собирање податоци

За целите на овој проект беа собрани податоци за продажбата од три различни извори:

- https://www.kaggle.com/datasets/yasserh/advertising-sales-dataset
- https://www.kaggle.com/datasets/yakhyojon/marketing-promotion
- https://huggingface.co/datasets/sitbayevalibek/Advertising-Sales-Dataset-Social-Network

Секој dataset вклучува различни променливи. Првиот dataset се состои од податоци за вложеност во реклами на телевизија, радио, весници и продажбата во милиони.

Вториот dataset содржи колони со податоци за сума потрошена за реклами на телевизија, радио, социјални медиуми, рекламирање преку инфлуенсери како и продажбите.

Третиот dataset е составен од податоци за потрошен буџет на реклами на Instagram, Facebook, YouTube, Twitter, Telegram, WhatsApp, TikTok, радио, ТВ, весници, веб сајтови и самата продажба во милиони.

2.2 Чистење и подготовка на податоци

Извршено е чистење на податоците за да се отстранат вредностите што недостасуваат, outliers и недоследностите.

Првото податочно множество немаше празни ќелии со податоци па немаше потреба од чистење и трансформирање на податоците освен отстранување на колони бидејќи во овој проект главна цел е поврзаноста меѓу рекламите на радио и ТВ и продажбата.

Вториот dataset се состоеше од недефинирани податоци т.е имаше празни полиња но бидејќи не стануваше збор за големо количество на вакви податоци истите беа отстранети.

```
df2 = pd.read_csv('/content/drive/MyDrive/Vnp project/Dummy Data HSS.csv')
```

```
[158] df2.head()
```

	TV	Radio	Social Media	Influencer	Sales
0	16.0	6.566231	2.907983	Mega	54.732757
1	13.0	9.237765	2.409567	Mega	46.677897
2	41.0	15.886446	2.913410	Mega	150.177829
3	83.0	30.020028	6.922304	Mega	298.246340
4	15.0	8.437408	1.405998	Micro	56.594181

Next steps: View recommended plots

```
[159] df2.isnull().sum()
```

```
TV              10
Radio            4
Social Media     6
Influencer       0
Sales            6
dtype: int64
```

```
[160] df2 = df2.dropna()

      print('Number of missing values:', df2.isnull().sum())
```

```
Number of missing values: TV             0
Radio          0
Social Media   0
Influencer     0
Sales          0
dtype: int64
```

Слика 1

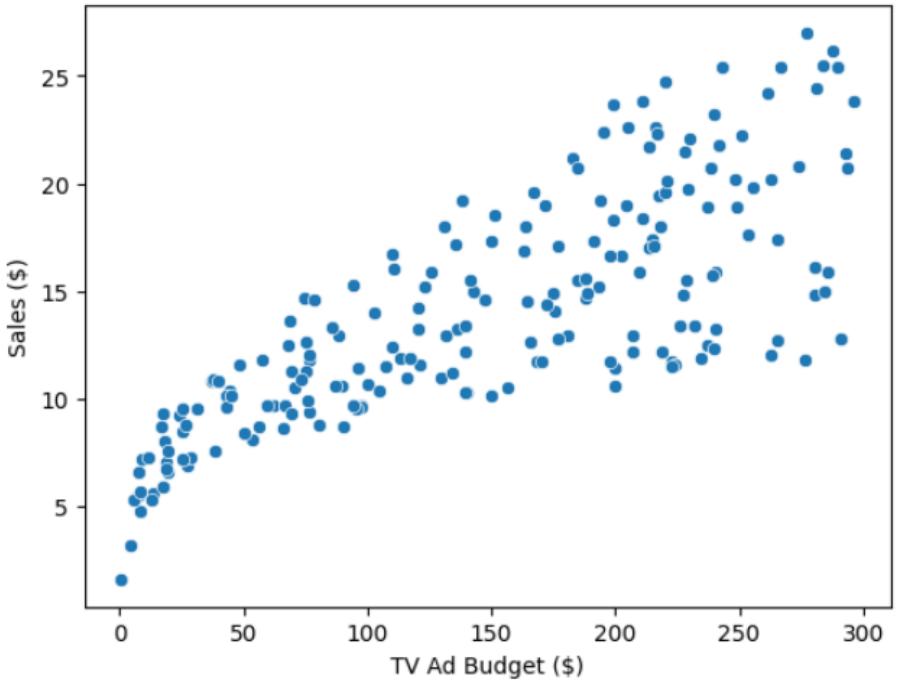
Аналогно на првиот dataet, третиот dataset немаше празни полиња па само беа отстранети колоните кои не се од важност за овој проект.

2.3 Истражувачка анализа на податоци (EDA)

Беа спроведени истражувачки анализи за да се разбере дистрибуцијата на продажбата, трендовите со текот на времето и корелациите помеѓу променливите. Техниките за визуелизација, како што се линиски графикони, хистограми и графики на расејување беа користени за истражување на податоците.

```
sns.scatterplot(df1, x = "TV Ad Budget ($)", y = "Sales ($)")
```

```
<Axes: xlabel='TV Ad Budget ($)', ylabel='Sales ($)'>
```

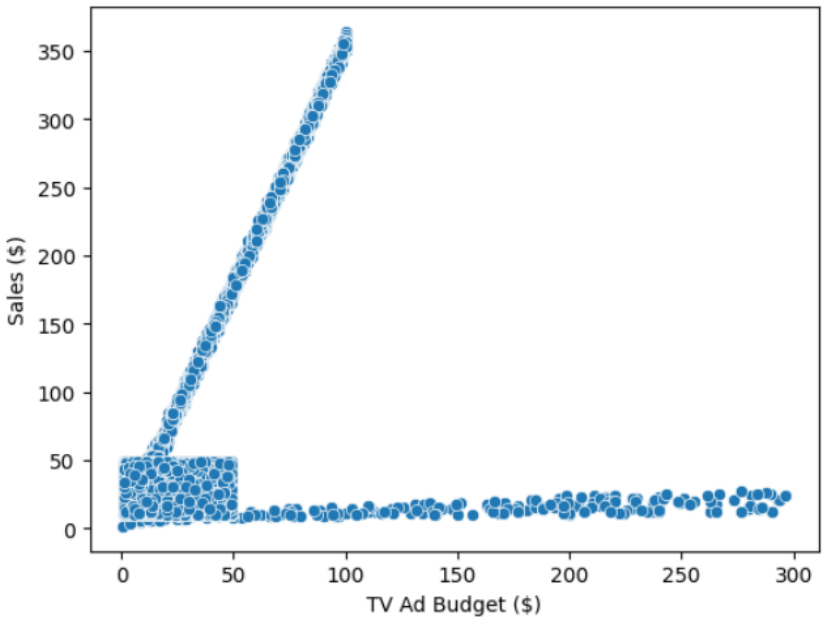


Слика 2

На Слика 2 е прикажана зависноста меѓу ТВ рекламен буџет и продажбата со што се забележува силна корелација.

```
sns.scatterplot(data, x = "TV Ad Budget ($)", y = "Sales ($)")
```

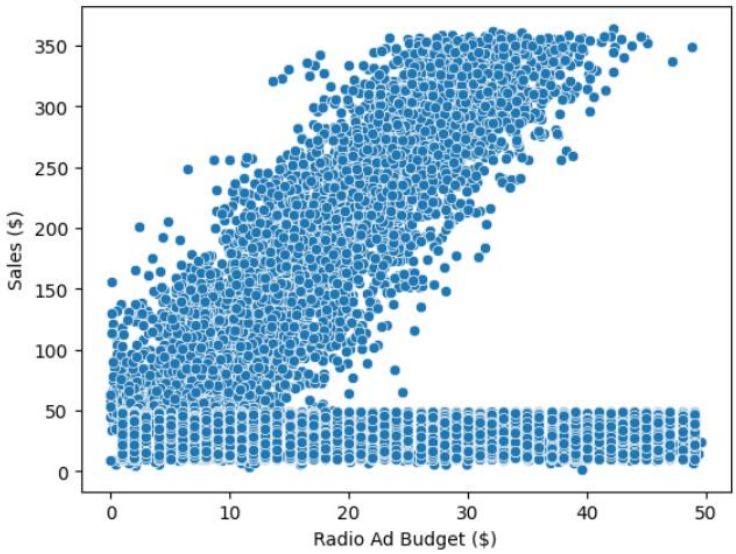
```
<Axes: xlabel='TV Ad Budget ($)', ylabel='Sales ($)'>
```



Слика 3

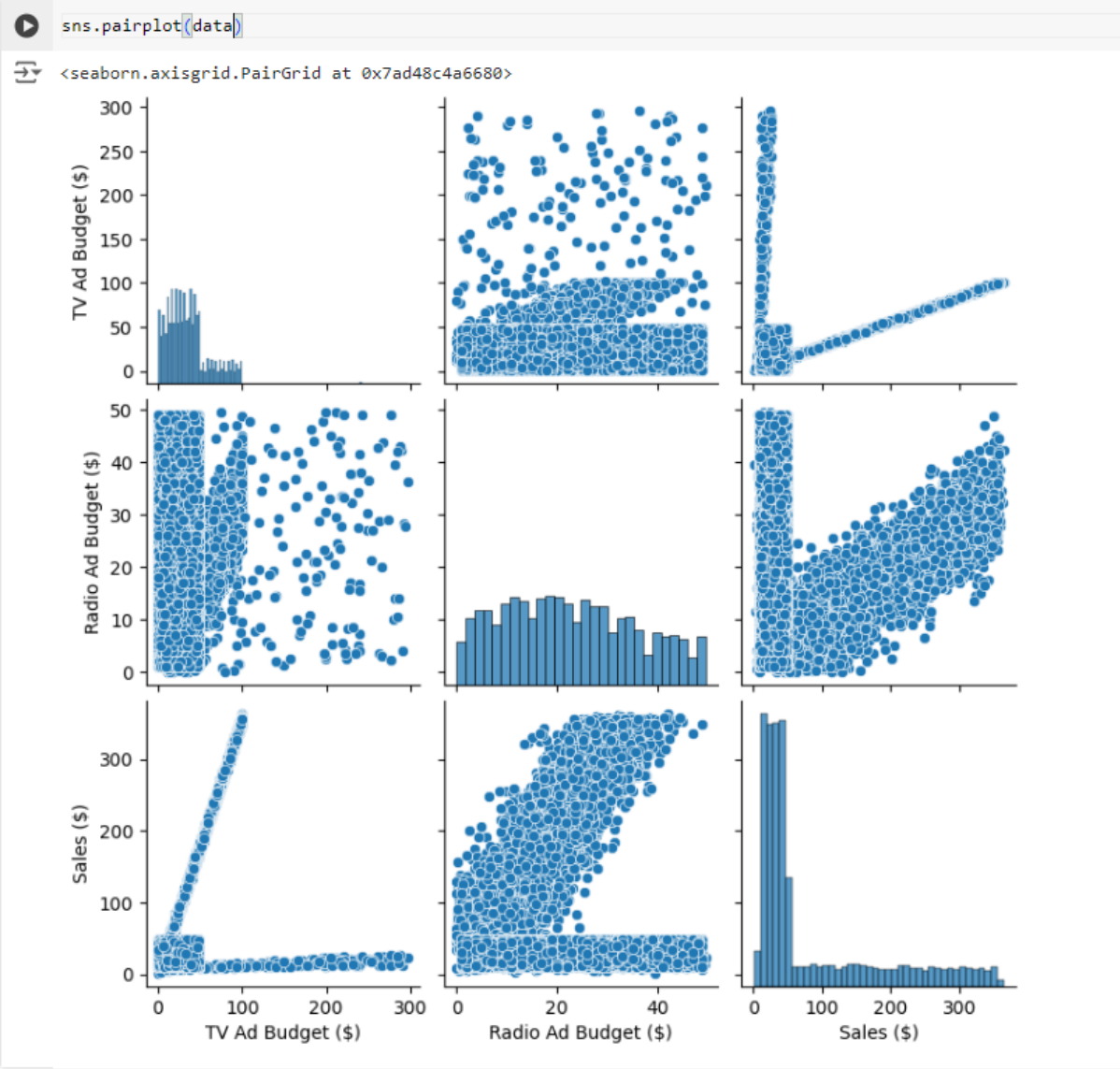
```
sns.scatterplot(data, x = "Radio Ad Budget ($)", y = "Sales ($)")
```

```
<Axes: xlabel='Radio Ad Budget ($)', ylabel='Sales ($)'>
```



Слика 4

Слика 3 и Слика 4 прикажува график на зависности на целото податочно множество од трите различни податочни извори.



Слика 5

2.4 Соединување на податоците

По извршените трансформации и манипулации со податоците од трите податочни множества, истите беа споени во едно податочно множество. Тоа множество е искористено за тренирање и предвидување по градењето на моделот.

2.5 Избор на модел

Врз основа на карактеристиките на податоците, беше избран соодветен модел за предвидување.

Избраниот модел за предвидување беше обучен на историски податоци, а неговите перформанси беа оценети со користење на стандардни метрики како што се Средна апсолутна грешка (MAE), Средна квадратна грешка (MSE) и Root Mean Squared Error (RMSE). Во овој случај станува збор за проблем на регресија бидејќи се врши предвидување на нумерички вредности и поради тоа за целта на овој проект беше искористен едноставен модел кој вклучува линеарна регресија.

```
data.head()
```

	TV Ad Budget ($)	Radio Ad Budget ($)	Sales ($)
0	16.0	6.566231	54.732757
1	13.0	9.237765	46.677897
2	41.0	15.886446	150.177829
3	83.0	30.020028	298.246340
4	15.0	8.437408	56.594181

Next steps: View recommended plots

```
[257] data.isnull().sum()
```

```
TV Ad Budget ($)       0
Radio Ad Budget ($)    0
Sales ($)              0
dtype: int64
```

```
[258] X = data[['TV Ad Budget ($)', 'Radio Ad Budget ($)']]
      Y = data['Sales ($)']
```

```
[259] model1= LinearRegression()
      X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=42)
      model1.fit(X_train, y_train)
      y_pred = model1.predict(X_test)
      r2_score_model1 = r2_score(y_test, y_pred)
      print("R-squared Score (without scaling):", r2_score_model2)
```

```
R-squared Score (without scaling): 0.8994380241009119
```

Слика 6

Резултати

Моделот за предвидување постигна r2 score од `0.8994380241009119`. Моделот може приближно точно да ги предвиди идните продажни трендови.

Анализата откри значајна позитивна корелација помеѓу расходите за рекламирање на ТВ и радио и обемот на продажба во сите три групи на податоци особено за рекламите на ТВ. При што вредностите MAE, MSE и RMSE укажуваат на силна предиктивна точност.

Дискусија

Интеграцијата на повеќе извори на податоци и сеопфатната анализа овозможија попрецизен модел за прогнозирање на продажбата. Со разгледување на податоци од повеќе извори, успеав да доловам поширок опсег на фактори кои влијаат на динамиката на продажбата, што доведе до подобрени предвидувања.

Заклучок

Како заклучок, предложениот пристап за прогнозирање на продажбата ја демонстрира ефективноста на интегрирањето на податоците од повеќе извори и спроведувањето сеопфатна анализа. Развиениот модел обезбедува вредни сознанија за процесите на деловно планирање и донесување одлуки, со потенцијални апликации во управувањето со залихите, распределбата на ресурсите и стратешкото планирање. Овој модел би овозможил компаниите да можат однапред да определуваат буџет за реклами на ТВ и радио.

Идни проекти

Фокусот на идините проекти ќе биде понатамошно усовршување на моделот за предвидување со инкорпорирање на дополнителни извори на податоци, истражување на алтернативни техники за моделирање и подобрување на приспособливоста и интерпретабилноста на моделот.

Користена литература

- https://www.kaggle.com/datasets/yasserh/advertising-sales-dataset
- https://www.kaggle.com/datasets/yakhyojon/marketing-promotion
- https://huggingface.co/datasets/sitbayevalibek/Advertising-Sales-Dataset-Social-Network