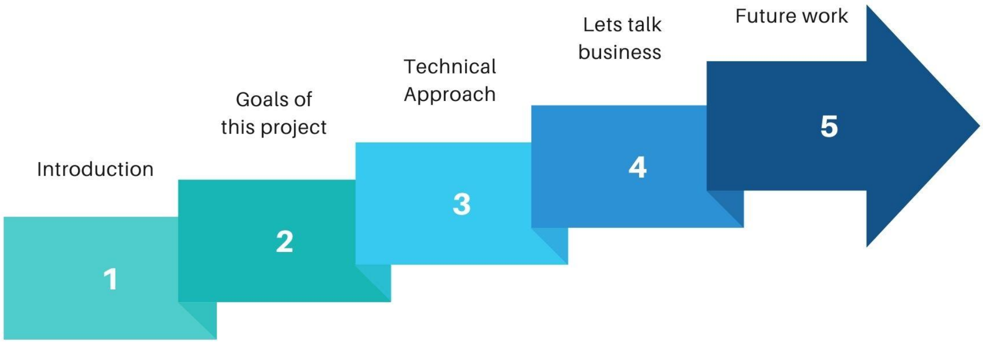
Report

OUR AGENDA FOR THE DAY

5-Step Ordering Process



1 Introduction

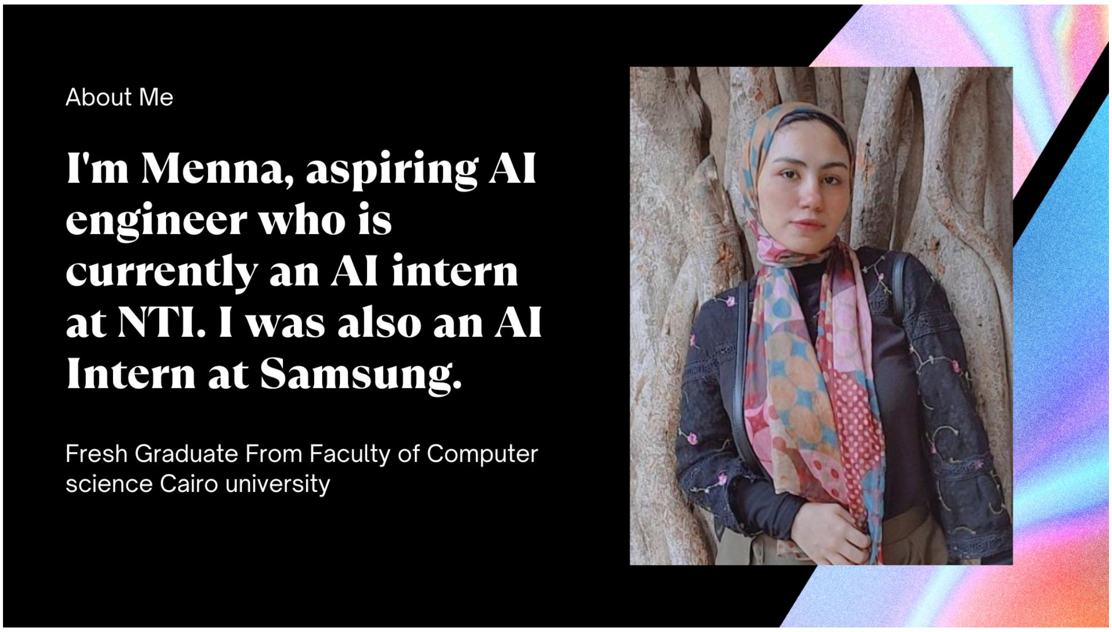
2 Goals of this project

3 Technical Approach

4 Lets talk business

5 Future work

Let me introduce myself :



About Me

I'm Menna, aspiring AI engineer who is currently an AI intern at NTI. I was also an AI Intern at Samsung.

Fresh Graduate From Faculty of Computer science Cairo university

The goals of This project:

To help the the management team to :

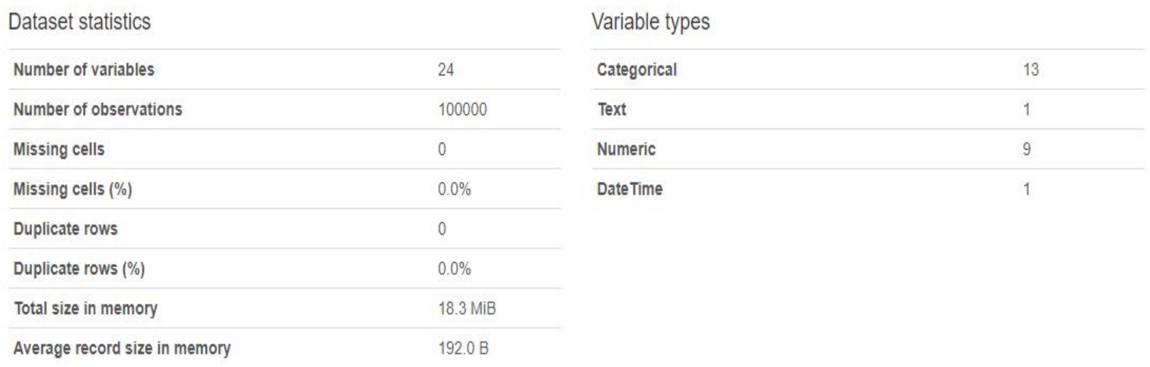
- See the value from using your platform for advanced analytics and the data platform challenges that accompany it.
- See if using advanced analytics may increase sales in general.

To help the the marketing team to :

- see if using advanced analytics you can help them increase their efficiency.

Technical Approach :

- Lets explore the data first:



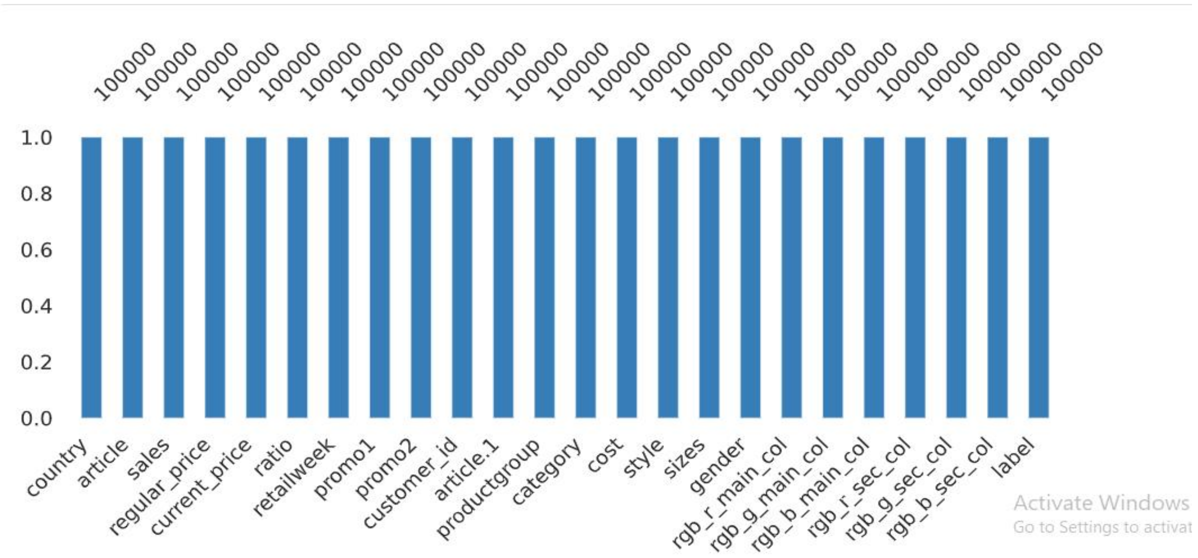
Dataset statistics

Number of variables	24
Number of observations	100000
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	18.3 MiB
Average record size in memory	192.0 B

Variable types

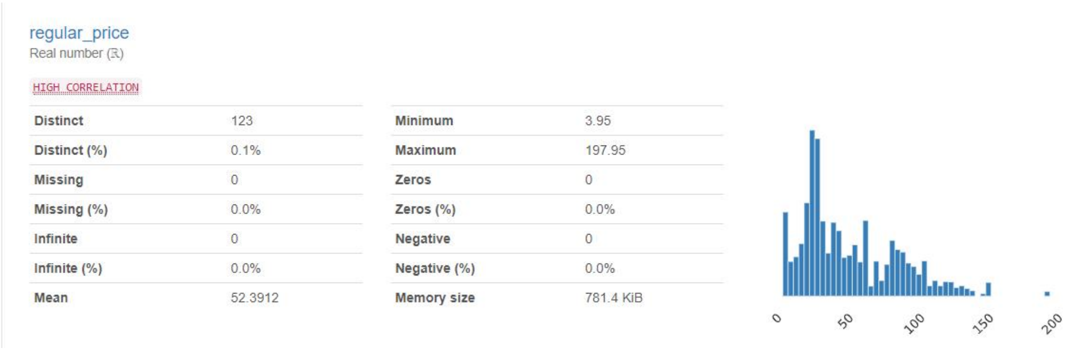
Categorical	13
Text	1
Numeric	9
DateTime	1

The data has No Missing Values :



The Most Important Features out of 24 :

-Regular_Price

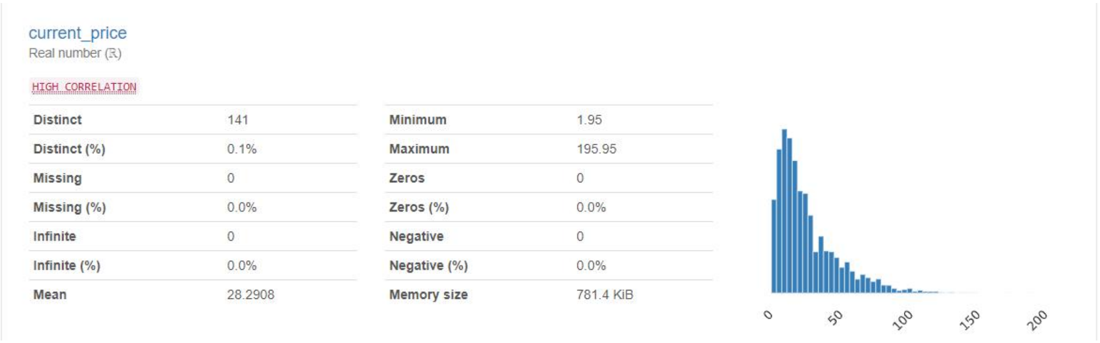


regular_price
Real number (ℝ)

HIGH CORRELATION

Distinct	123	Minimum	3.95
Distinct (%)	0.1%	Maximum	197.95
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	52.3912	Memory size	781.4 KiB

- Current_Price

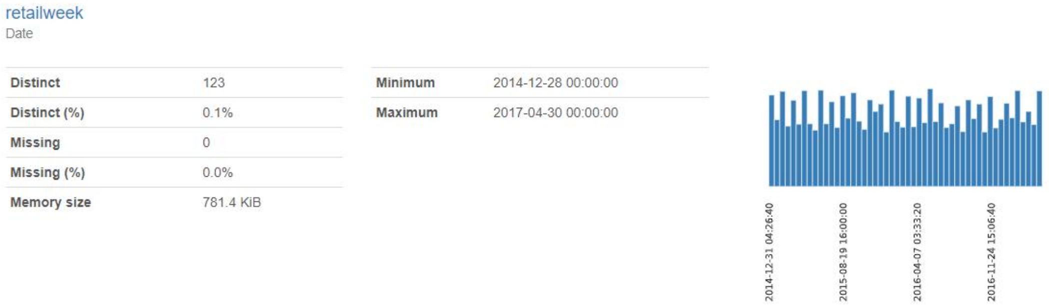


current_price
Real number (ℝ)

HIGH CORRELATION

Distinct	141	Minimum	1.95
Distinct (%)	0.1%	Maximum	195.95
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	28.2908	Memory size	781.4 KiB

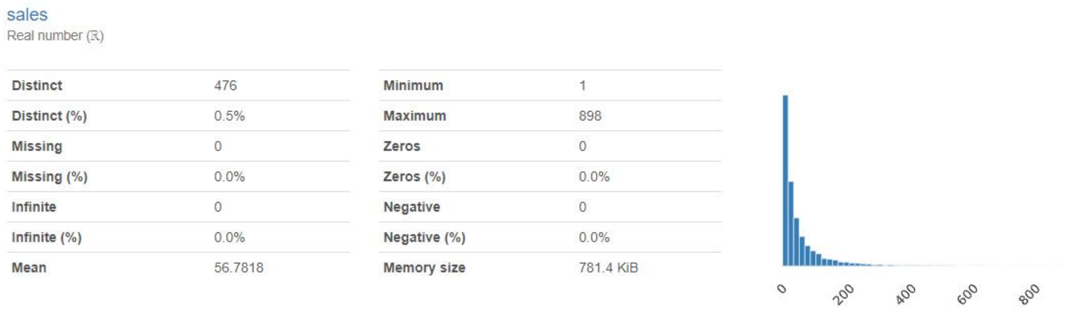
-Retailweek



retailweek
Date

Distinct	123	Minimum	2014-12-28 00:00:00
Distinct (%)	0.1%	Maximum	2017-04-30 00:00:00
Missing	0		
Missing (%)	0.0%		
Memory size	781.4 KiB		

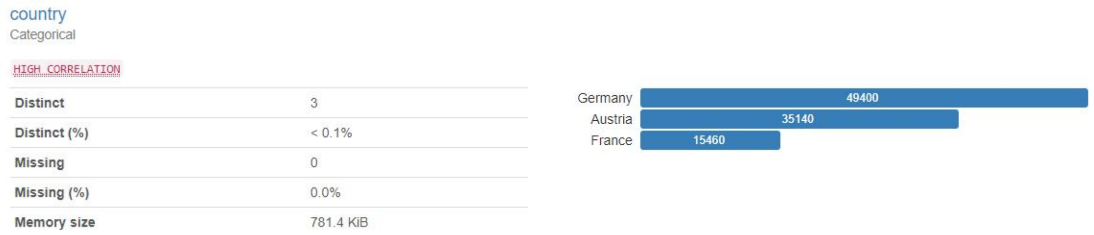
-Sales



sales
Real number (ℝ)

Distinct	476	Minimum	1
Distinct (%)	0.5%	Maximum	898
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	56.7818	Memory size	781.4 KiB

- Country



country
Categorical

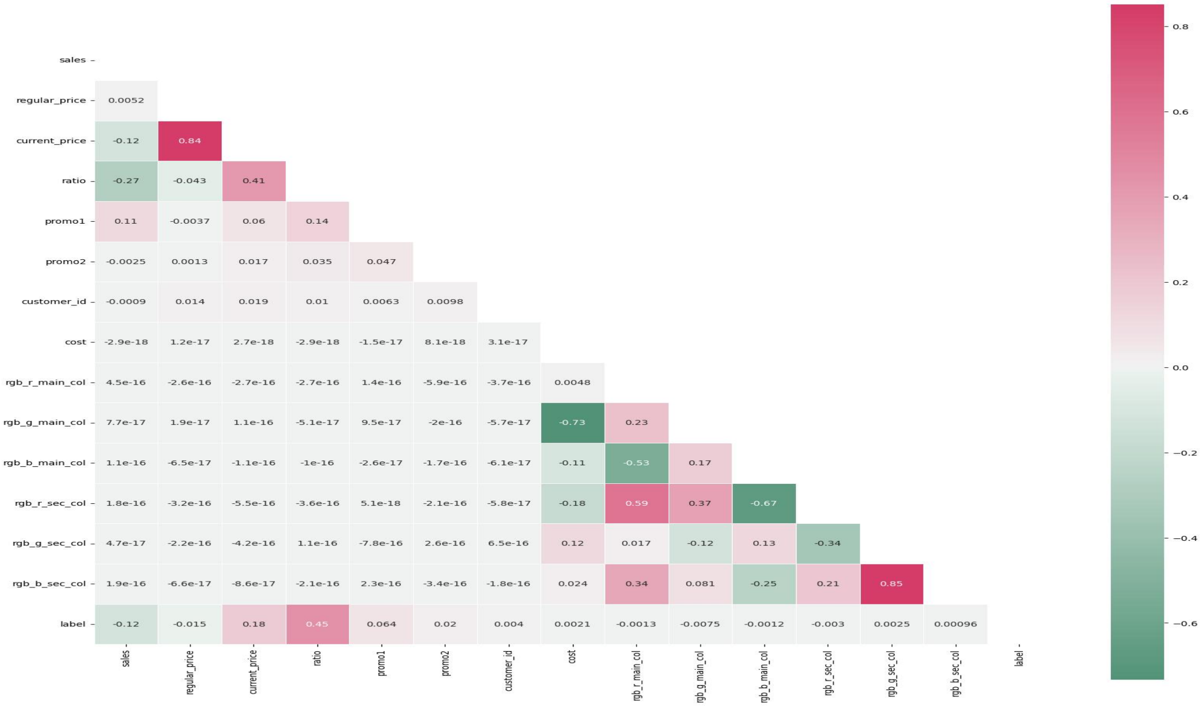
HIGH CORRELATION

Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	781.4 KiB

Lets dive into Our Approach:

- Data visualization
- Analytical insights with EDA
- Preprocessing
- Modeling

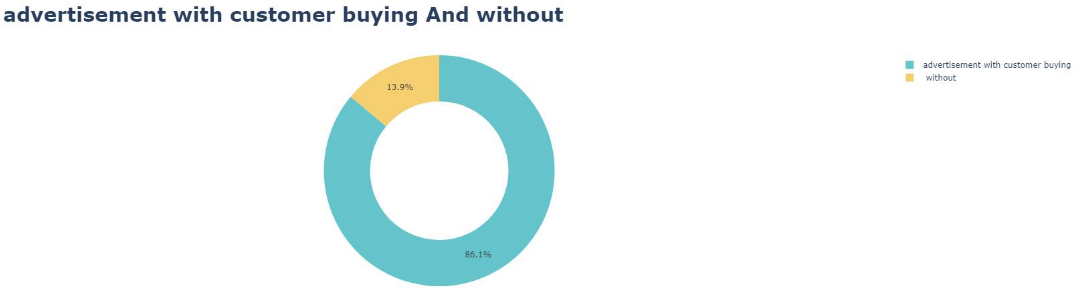
Data visualization :



	sales	regular_price	current_price	ratio	promo1	promo2	customer_id	cost	rgb_r_main_col	rgb_g_main_col	rgb_b_main_col	rgb_r_sec_col	rgb_g_sec_col	rgb_b_sec_col	label
sales															
regular_price	0.0052														
current_price	-0.12	0.84													
ratio	-0.27	-0.043	0.41												
promo1	0.11	-0.0037	0.06	0.14											
promo2	-0.0025	0.0013	0.017	0.035	0.047										
customer_id	-0.0009	0.014	0.019	0.01	0.0063	0.0098									
cost	-2.9e-18	1.2e-17	2.7e-18	-2.9e-18	-1.5e-17	8.1e-18	3.1e-17								
rgb_r_main_col	4.5e-16	-2.6e-16	-2.7e-16	-2.7e-16	1.4e-16	-5.9e-16	-3.7e-16	0.0048							
rgb_g_main_col	7.7e-17	1.9e-17	1.1e-16	-5.1e-17	9.5e-17	-2e-16	-5.7e-17	-0.73	0.23						
rgb_b_main_col	1.1e-16	-6.5e-17	-1.1e-16	-1e-16	-2.6e-17	-1.7e-16	-6.1e-17	-0.11	-0.53	0.17					
rgb_r_sec_col	1.8e-16	-3.2e-16	-5.5e-16	-3.6e-16	5.1e-18	-2.1e-16	-5.8e-17	-0.18	0.59	0.37	-0.67				
rgb_g_sec_col	4.7e-17	-2.2e-16	-4.2e-16	1.1e-16	-7.8e-16	2.6e-16	6.5e-16	0.12	0.017	-0.12	0.13	-0.34			
rgb_b_sec_col	1.9e-16	-6.6e-17	-8.6e-17	-2.1e-16	2.3e-16	-3.4e-16	-1.8e-16	0.024	0.34	0.081	-0.25	0.21	0.85		
label	-0.12	-0.015	0.18	0.45	0.064	0.02	0.004	0.0021	-0.0013	-0.0075	-0.0012	-0.003	0.0025	0.00096	

Analytical insights with EDA :

-How many people Buy our Products ?

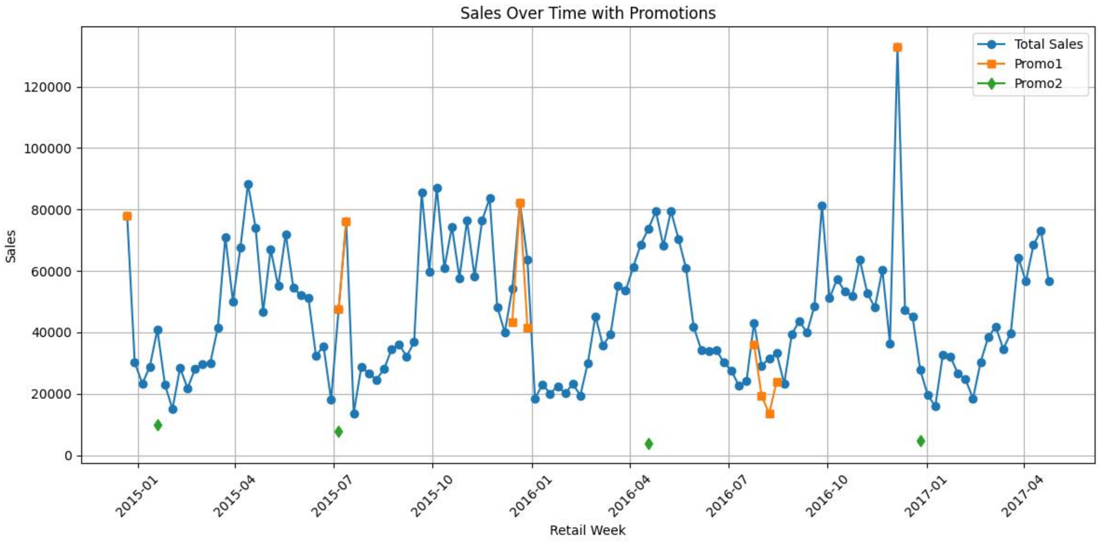


From a Technical perspective The label is imbalanced as it has

- 13.9% of people dont do advertising
- 86.1 do advertising

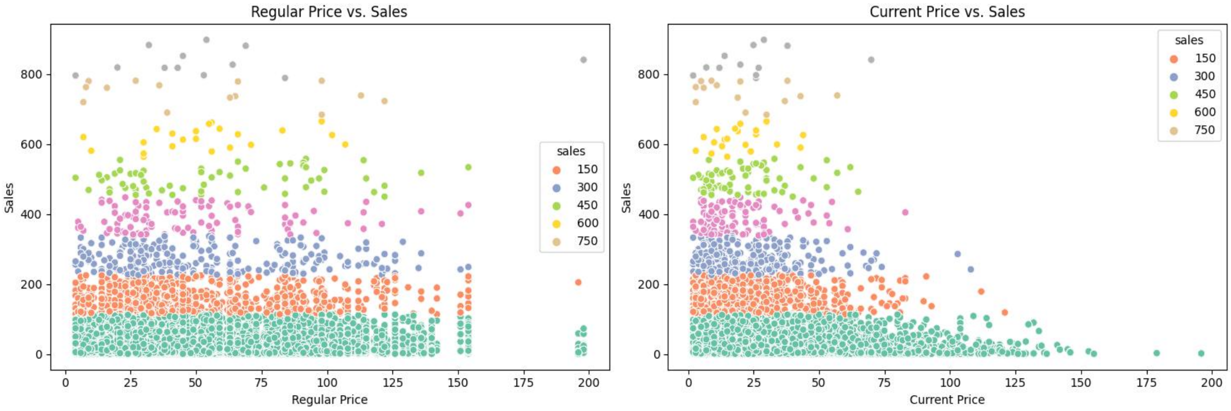
From Sales And marketing perspective its good that the highest percentage of people do advertising

-What do you think the Impact of sales Over Time with Promotions ?



- Sales increase during promotional periods so it's a positive indicator of the effectiveness of our promotional strategies

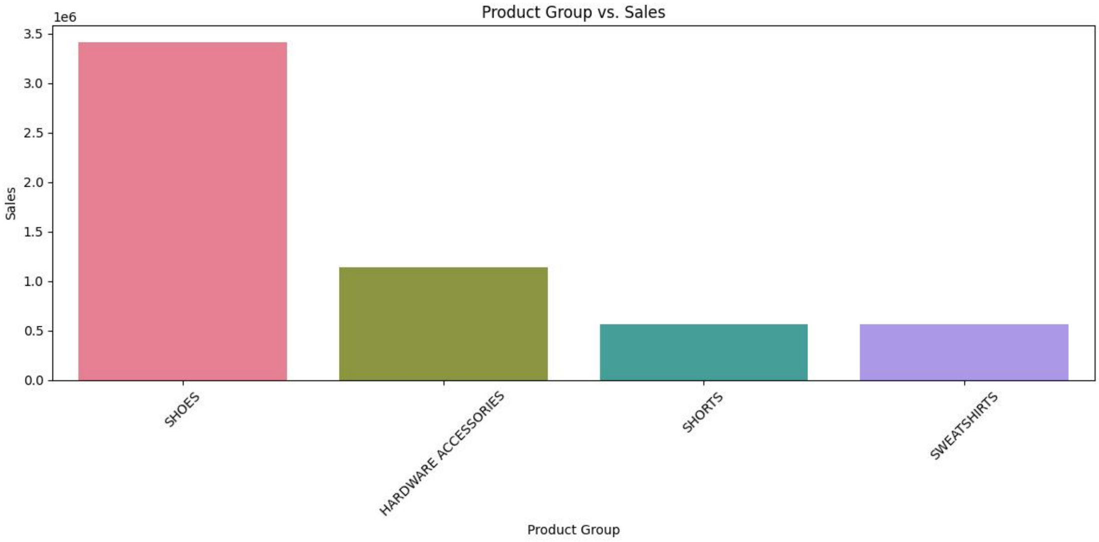
How changes in regular and current prices affect sales ?



They have the same effect on sales and it increase with both of them and that means :

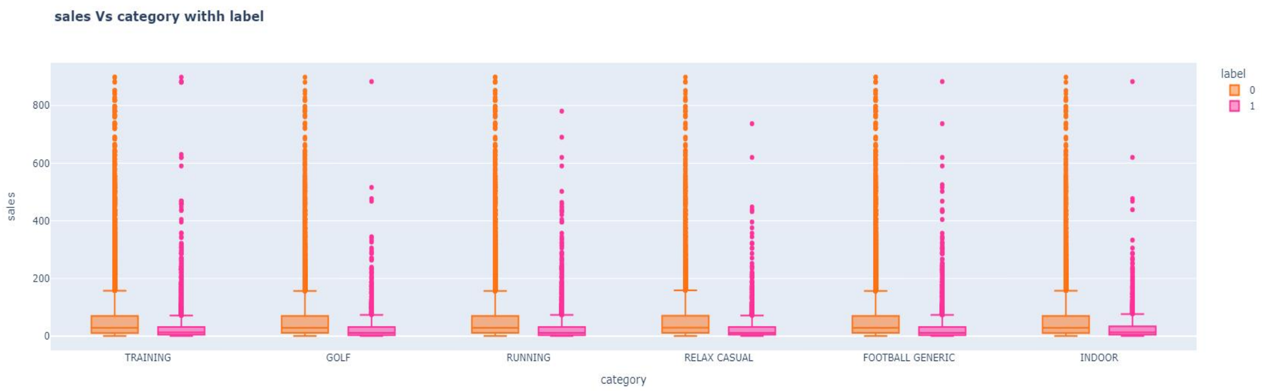
- there is a strong demand for the product
- Customers might be willing to pay more for the product due to its popularity and unique features.

The Most Popular Product :



- As we see the most popular Product that customers are interested in purchasing is Shoes

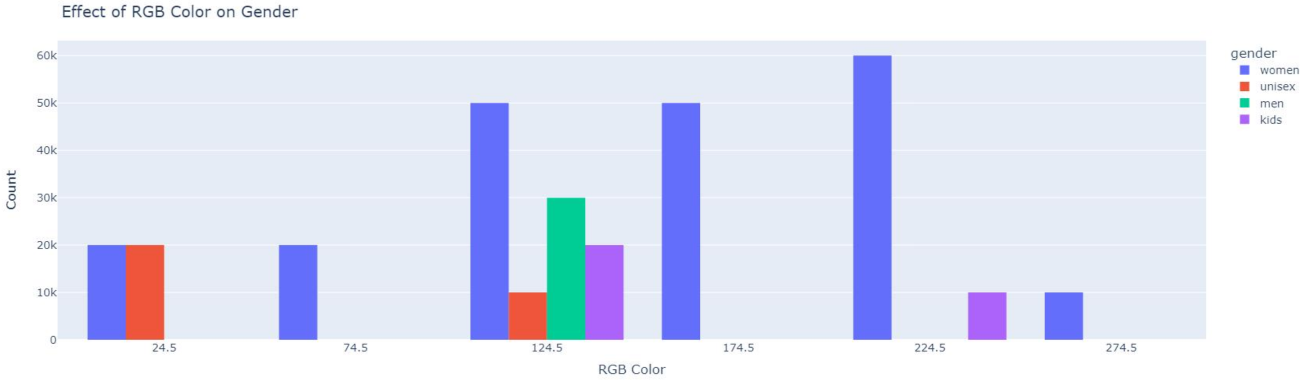
How sales vary across different product categories ?

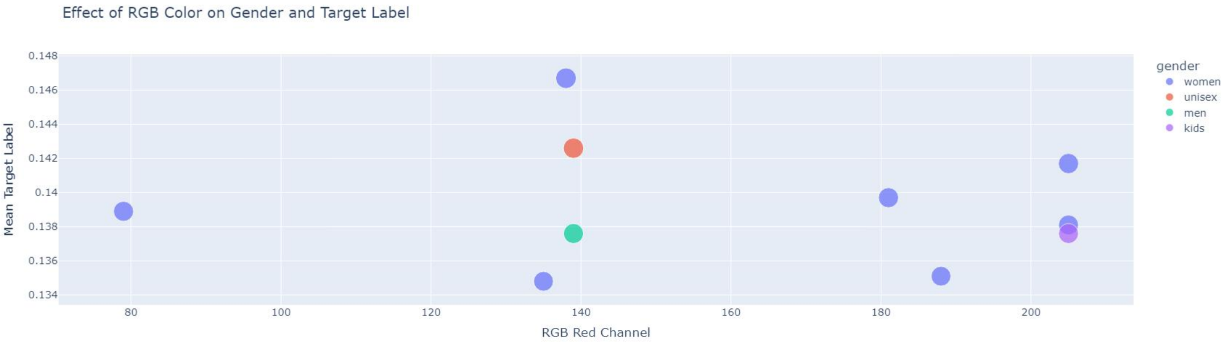


Sales Don't Vary Much that mean :

- **Customers are purchasing products from various categories, indicating a balanced demand but the customers still dont buy that much as we said the data is imbalanced in the advertising process**

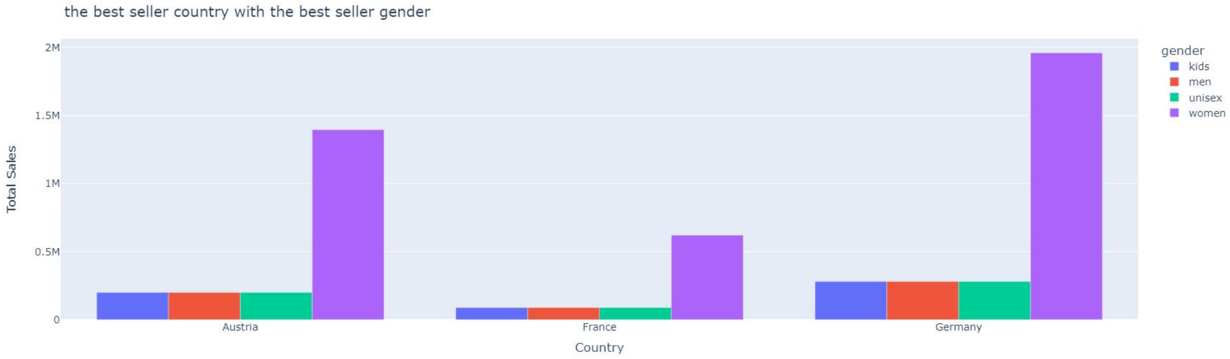
Which Gender interested to buy the product depending on colors ?





Women has the most interest in Colors and the product has more popularity with women

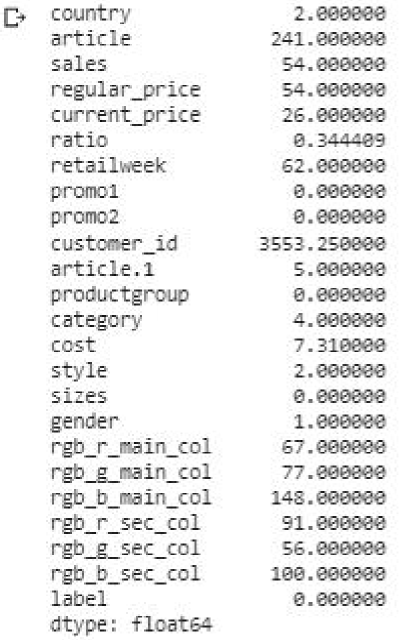
Which country is the best Buyer country with the best Buyer gender:



The best buyer country is Germany and women is the best buyer Gender

Preprocessing :

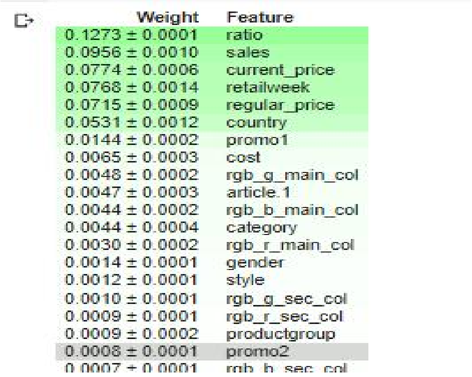
- **Discovering Outliers :**



```
country              2.000000
article            241.000000
sales               54.000000
regular_price       54.000000
current_price       26.000000
ratio                0.344409
retailweek          62.000000
promo1               0.000000
promo2               0.000000
customer_id       3553.250000
article.1            5.000000
productgroup         0.000000
category             4.000000
cost                 7.310000
style                2.000000
sizes                0.000000
gender               1.000000
rgb_r_main_col      67.000000
rgb_g_main_col      77.000000
rgb_b_main_col     148.000000
rgb_r_sec_col       91.000000
rgb_g_sec_col       56.000000
rgb_b_sec_col      100.000000
label                0.000000
dtype: float64
```

- **Dropping 2 columns :** **['article','customer_id']**

-Discovering Important Features:



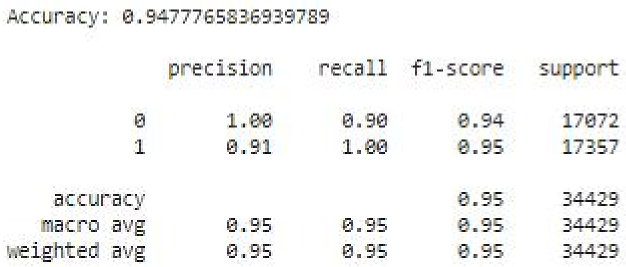
Weight	Feature
0.1273 ± 0.0001	ratio
0.0956 ± 0.0010	sales
0.0774 ± 0.0006	current_price
0.0768 ± 0.0014	retailweek
0.0715 ± 0.0009	regular_price
0.0531 ± 0.0012	country
0.0144 ± 0.0002	promo1
0.0065 ± 0.0003	cost
0.0048 ± 0.0002	rgb_g_main_col
0.0047 ± 0.0003	article.1
0.0044 ± 0.0002	rgb_b_main_col
0.0044 ± 0.0004	category
0.0030 ± 0.0002	rgb_r_main_col
0.0014 ± 0.0001	gender
0.0012 ± 0.0001	style
0.0010 ± 0.0001	rgb_g_sec_col
0.0009 ± 0.0001	rgb_r_sec_col
0.0009 ± 0.0002	productgroup
0.0008 ± 0.0001	promo2
0.0007 ± 0.0001	rgb_b_sec_col

-Random Oversampling to balance our label which made the accuracy more higher

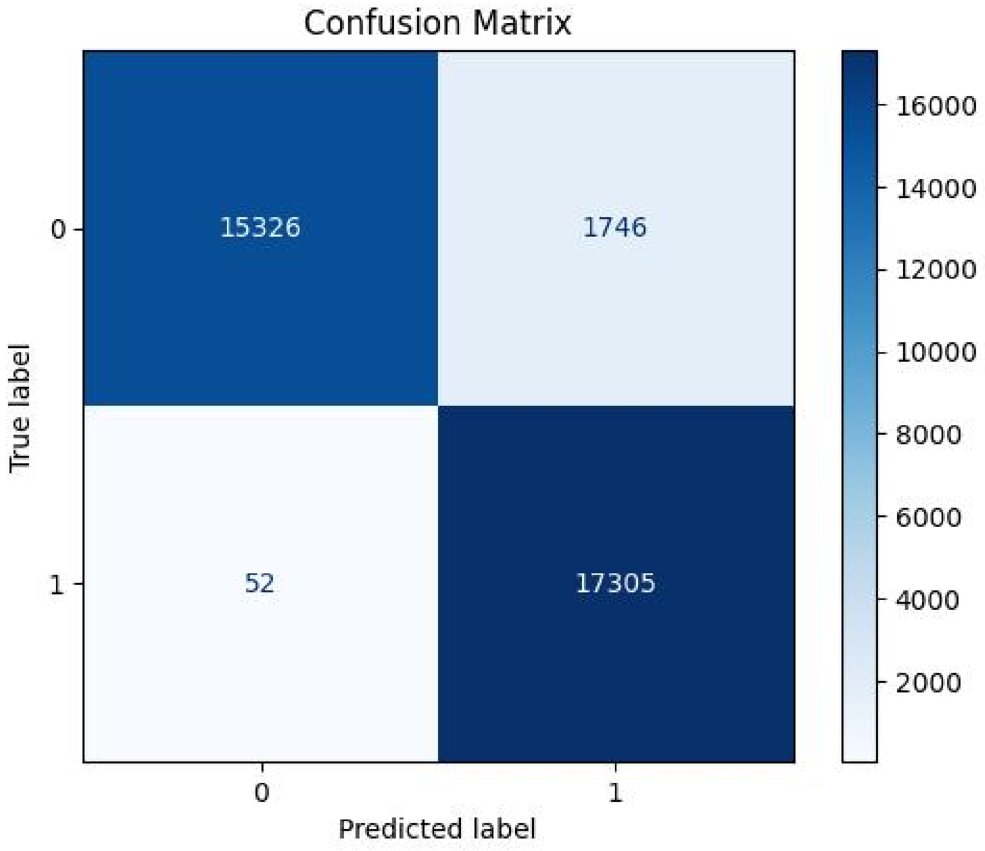
Modeling :

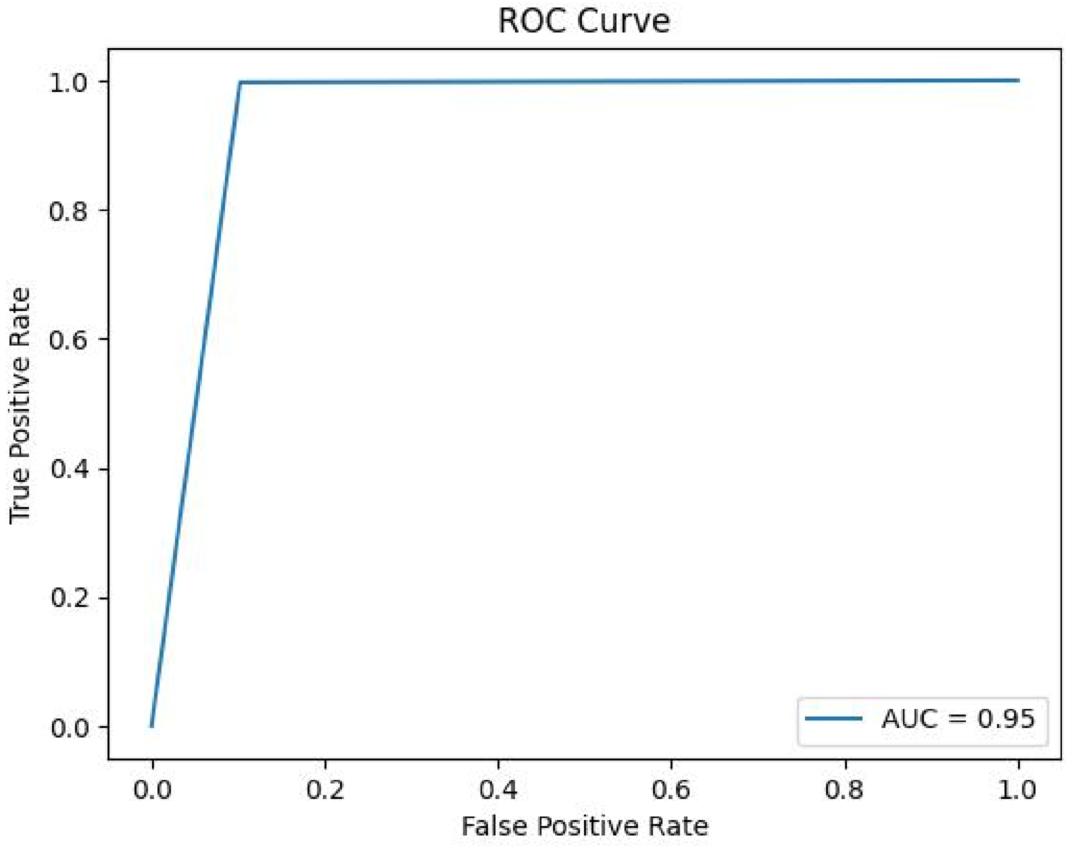
I tried Logistic Regression,Decision Tree,KNN,XGBOOST
Random Forest and xGBoost were the best Models so we would go with any of them
I chose these Models because they are suitable for Classification Problem like our Problem

Random Forest :

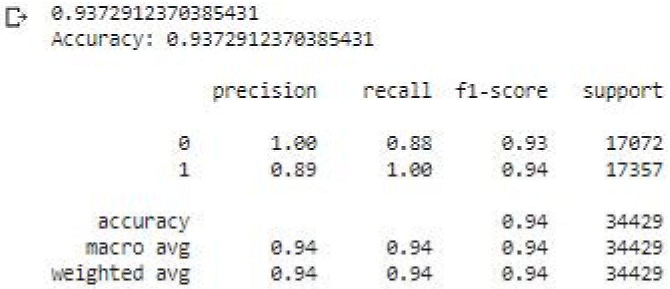


```
Accuracy: 0.9477765836939789
              precision    recall  f1-score   support

           0       1.00      0.90      0.94     17072
           1       0.91      1.00      0.95     17357

    accuracy                           0.95     34429
   macro avg       0.95      0.95      0.95     34429
weighted avg       0.95      0.95      0.95     34429
```

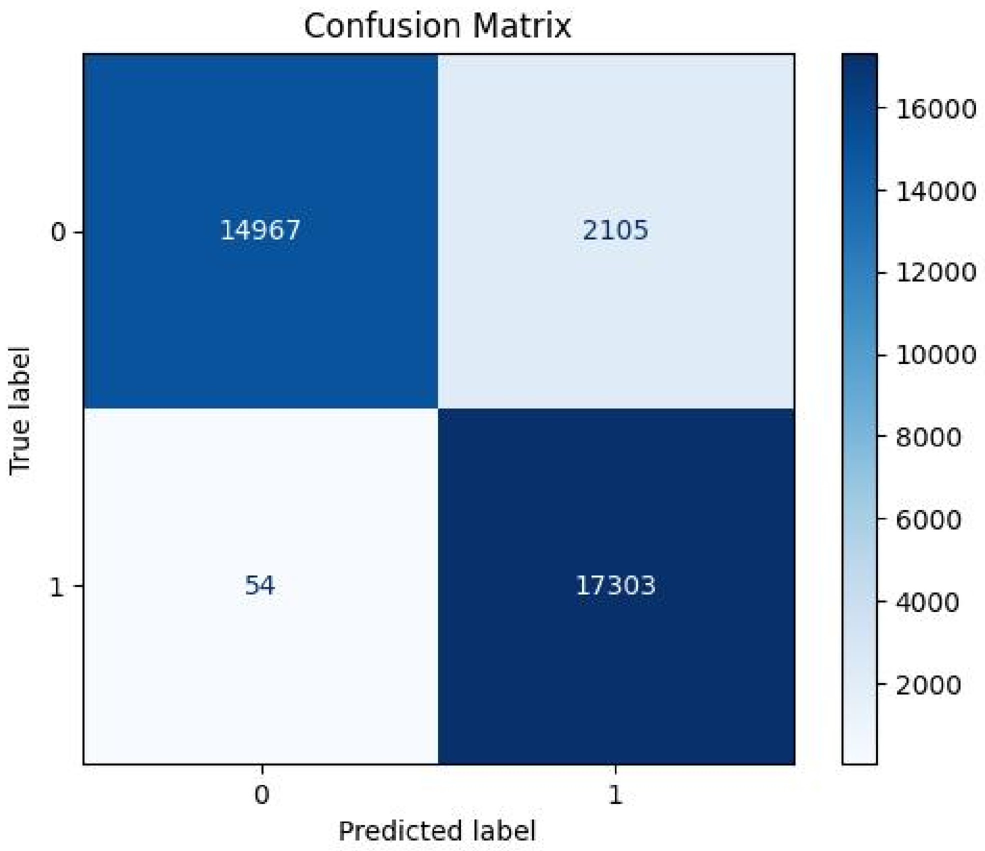


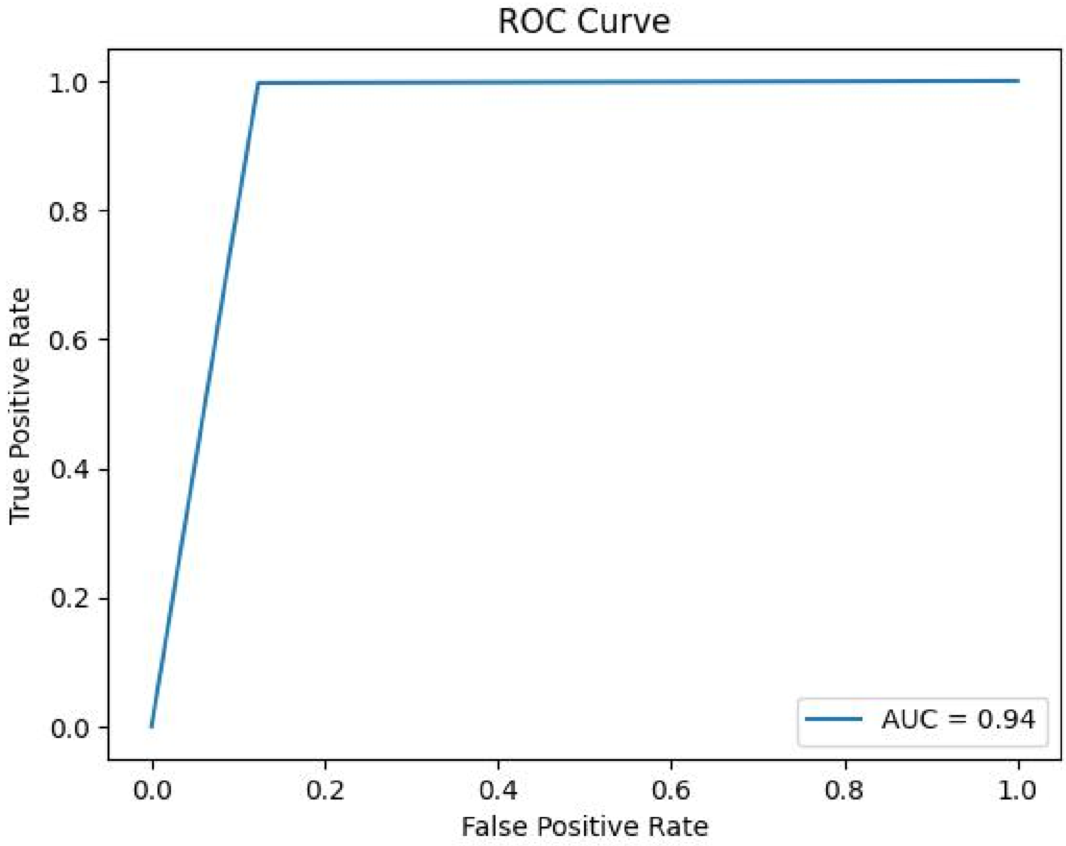


-XGBoost:

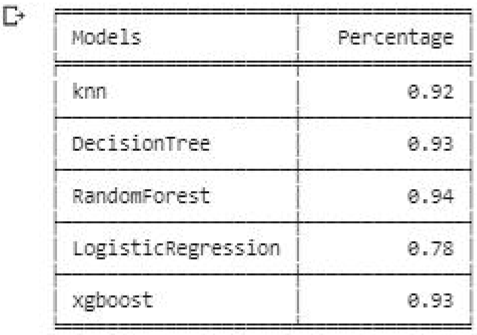


```
0.9372912370385431
Accuracy: 0.9372912370385431

              precision    recall  f1-score   support

           0       1.00      0.88      0.93     17072
           1       0.89      1.00      0.94     17357

    accuracy                           0.94     34429
   macro avg       0.94      0.94      0.94     34429
weighted avg       0.94      0.94      0.94     34429
```


Confusion Matrix
True label
Predicted label
0
1
0
1
14967
2105
54
17303
16000
14000
12000
10000
8000
6000
4000
2000


ROC Curve
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.94

-The accuracy of the different Models we Tried :



```
╒════════════════════╤══════════════╕
│ Models             │   Percentage │
╞════════════════════╪══════════════╡
│ knn                │         0.92 │
├────────────────────┼──────────────┤
│ DecisionTree       │         0.93 │
├────────────────────┼──────────────┤
│ RandomForest       │         0.94 │
├────────────────────┼──────────────┤
│ LogisticRegression │         0.78 │
├────────────────────┼──────────────┤
│ xgboost            │         0.93 │
╘════════════════════╧══════════════╛
```

Future work :
We would like to work more on the products to be more good and popoular among different people