****

NTI National Telecommunication Institute

Sport wear group marketing analysis and modeling

**By: Mohammed Hussein**

**July 2021**

**Data understanding:**

The data is from marketing campaign for sport wear group; this data contains info about the product, customer and marketing campaign. The label is the convergence of a customer 1 if a specific customer buy this product and 0 if not.

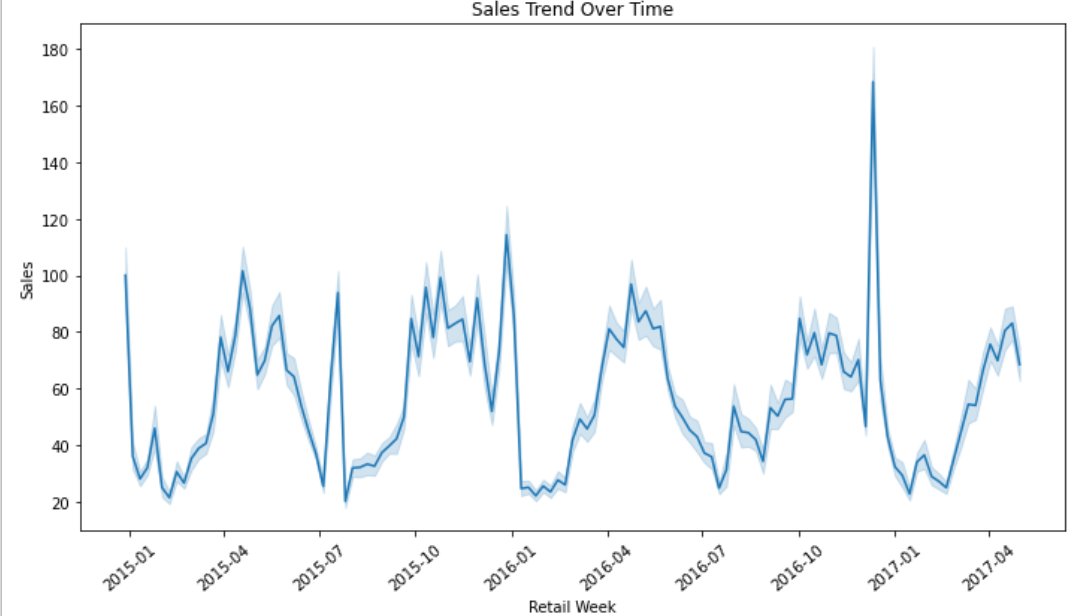
This data is 100000 row and 24 columns (100000 \* 24).

This data is collected from 4/1/2015 to 30/4/2017.

**Data cleaning and EDA:**

There is no null values or duplicates in data.

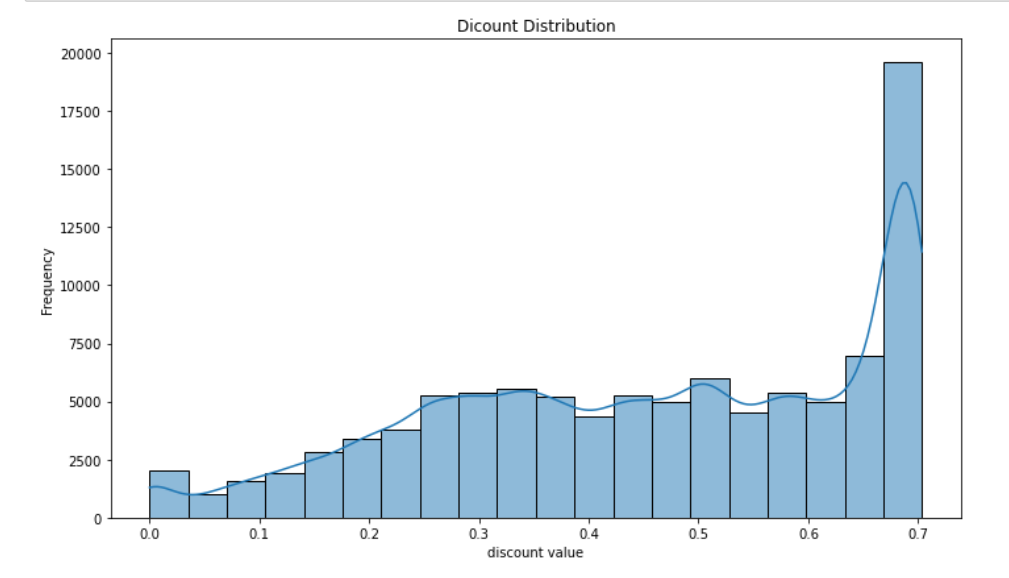
the sales changes over time but we can observe from the graph that sales increasing from march to May in every year and then starts to decreases until august, and then increasing from august to December and then starts to decrease until may come again.



So, we can draw an insight from this graph to increase marketing through those month.

We can engineer new features from the data:

* engineer feature revenue from sales and price features
* engineer the feature discount from ration



The most used discount is 70 %.

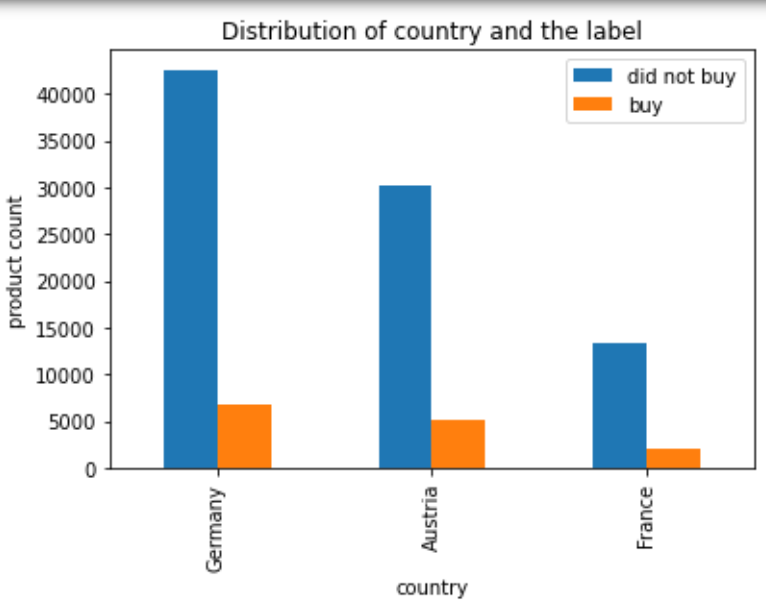
The dominate product group is shoes.

The dominate category is training, running and football generic.

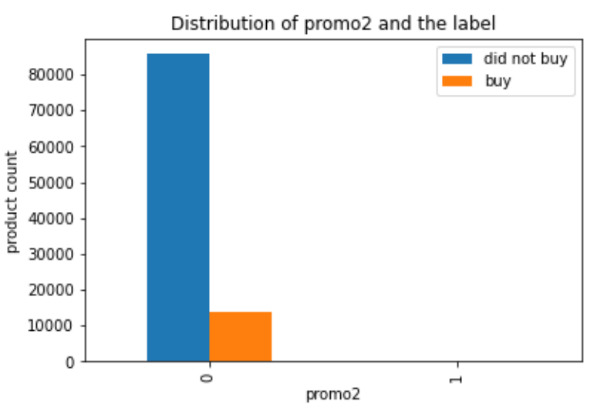
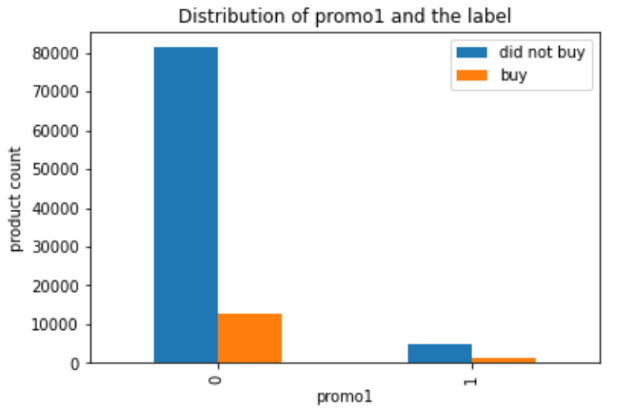
The target consumer that dominate the marketing data is women gender

The regular style is the most dominate style.

Most of sales value are between 20 and 100.



The marketing campaign is Germany most focused but the convergence of customer is the same for all countries.



Promotions effect is not clear in the graph but we represent it numerically to draw insights.

Promotion 1 (Media ads):

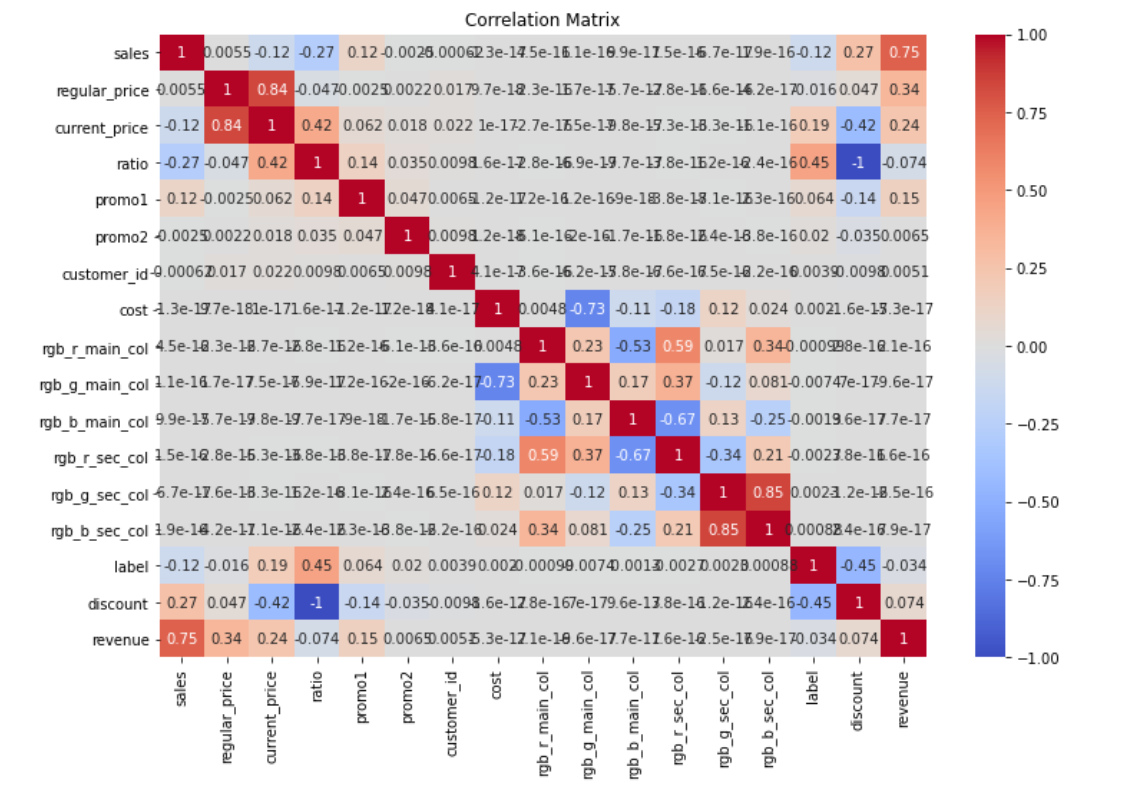
* impact on sales 77 %
* impact on label 69 %

Promotion 2 (Store events):

* impact on sales -5 %
* impact on label 72 %

It's interesting to note that while both promotions seem to have a positive impact on purchase decisions (increasing the likelihood of conversion), their impacts on sales differ. Promotion 1 (media advertisement) had a substantial positive impact on both sales and purchase decisions, Promotion 2 (store events), however, had a negative impact on sales but still managed to influence purchase decisions positively. The contrasting impacts could be due to various factors such as the effectiveness of media advertisement, the nature of store events, customer preferences, and the types of products promoted.

Correlation matrix:



Revenue and sales is highly correlated.

Ratio and revenue are the same impact so we can remove one of them.

Blue and green secondary are highly correlated so we can remove one of them.

Green main highly corr with the cost we can remove one of them.

**Data Preprocessing:**

* Label encoding for **sizes** and **style,** as they are ordinal features.
* One hot encoding for the nominal categorical features we have
* Nominal features = ['country', 'productgroup', 'category','sizes','gender'].
* Split retailweek into 3 columns ( year, month ,day).
* drop day and year as the only features has impact on sales is month as visualized in the time series graph

**Machine learning model before resampling:**

Training Model LR

--------------

Training Accuracy: 0.8605344965904532

Testing Accuracy: 0.8601584436421982

Training Model DT

--------------

Training Accuracy: 0.9987590252707581

Testing Accuracy: 0.7979843561973526

Training Model RF

--------------

Training Accuracy: 0.998671279582832

Testing Accuracy: 0.825812274368231

Training Model XGB

--------------

Training Accuracy: 0.8649468511833133

Testing Accuracy: 0.858754512635379

Training Model Naive Bayes

--------------

Training Accuracy: 0.8607977336542318

Testing Accuracy: 0.8606097071801043

**Apply oversampling:**

As shown in the plot below our data is imbalanced with more than 80,000 within label 0 and less than 20,000 with label 1.

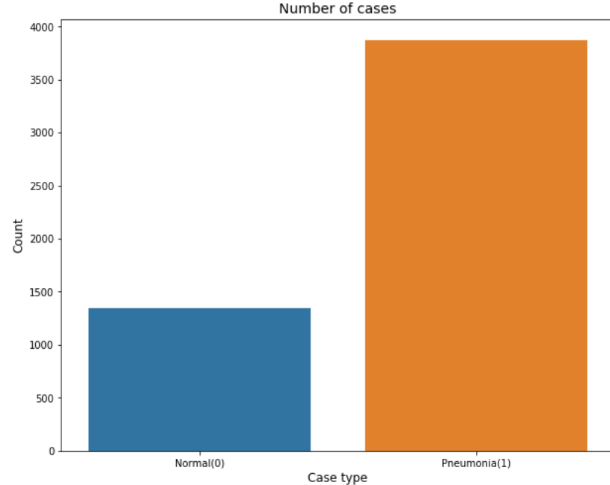


Figure 1 before oversampling

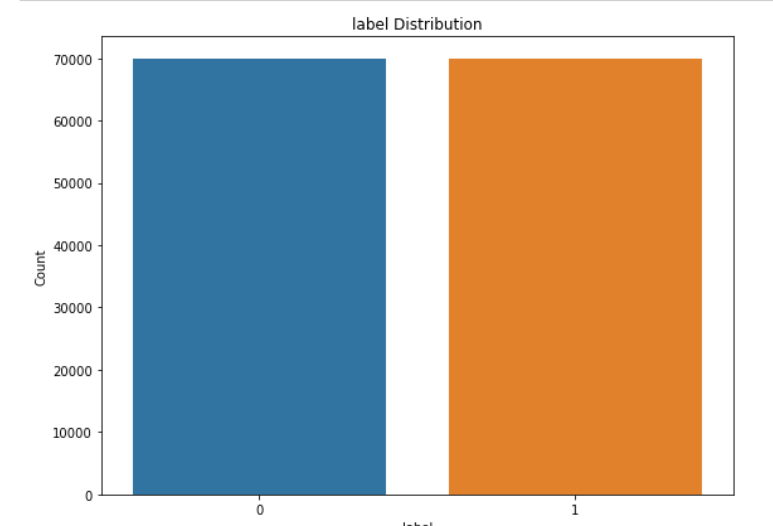


Figure 2 after oversampling

**Training again after resampling:**

Training result after resampling gives bad result except for decision tree and random forest.

**Best model:**

Decision tree gives **95 % accuracy** on test data.

Recall Score: 0.9510357142857143

Precision Score: 0.9510357142857143[¶](http://localhost:8888/notebooks/sport%20wear%20use%20case.ipynb#Precision-Score-:--0.9510357142857143)

Testing F-1: 0.9510357142857143

Testing F-Beta: 0.9510357142857144