# Textile Chain Retail Analysis and Customer Segmentation

---

**Contributor,**
**Rajathadri A S**

## Introduction:

The aim of this project was to analyze a large dataset of a textile retail store containing 1.1 million rows and 10 columns, and to segment the customers into different categories based on their purchase behavior. The data contains information about customer transactions and various attributes related to each transaction. The company wants to analyze this data to gain insights into customer behavior and preferences in order to improve the store's marketing and sales strategies. The project involves data cleaning, data exploration, feature engineering, and data visualization techniques to gain insights into customer behavior and preferences. The ultimate goal is to provide actionable insights to the store that will help them improve their sales and marketing strategies and enhance customer satisfaction. The project was carried out for Capillary Technologies.

## Data Wrangling:

The dataset contained numerous duplicate entries, so a new quantity feature was created that counted the number of duplicates and incremented the quantity value for every duplicate found. However, iterating through 1.1 million rows was difficult due to the high memory requirements and time constraints. Hence, only a part of the dataset was considered for further analysis. After this, data wrangling was carried out which included converting features to relevant data types, handling missing and invalid values, converting textual data to lowercase, and performing RFM segmentation to classify customers into ranks based on their purchase behavior.

## Exploratory Data Analysis:

Univariate and bivariate analysis were carried out to understand the relationships between features. It was found that the majority of the discounts were offered only to items of size 38, the majority of sales were generated from the stores in the 'south zone', who distributed discounts equally among all items and sizes, and our customer segmentation had only 4.2% of the customers in the Top-customer category. Based on these insights,

strategic recommendations were made for stores in every zone based on their sales performance.

## Machine Learning:

Categorical encoding was carried out for required categorical variables, and three machine learning models were built - Logistic Regression, Random Forest, and XGB Classifier models. These models were trained on the dataset to predict customer behavior and classify them into different categories.

## Recommendations:

After performing Exploratory Data Analysis, we have decided to make strategic recommendations that are distinctive to individual zones. There are 6 zones and our recommendations for each zone are as shown below.

*South Zone:*

Although South zone has the highest number of sales, these sales are mainly obtained through Low, Medium and Lost customers. It is not an ideal scenario for any sales platform in the long run. If these customers are not encouraged to buy repeatedly from the stores, we can lose these customers permanently.

*East Zone :*

It seems like the east zone stores are slightly new since they do not have a significant number of sales. These stores should focus on overall marketing of their stores without focusing on the customer segment to provide offers and discounts to all customers.

North Zone :

These stores have decent sales numbers, but they need to focus more on Low and Lost Customers. Customers in the low segment need to be offered with coupon codes to encourage them to be repeat customers, whereas the lost customers should be offered with promotional offers in order to re-acquire these customers.

*West Zone :*

These stores have decent sales numbers, but they need to focus more on Low and Lost Customers. Customers in the low segment need to be offered with coupon codes to encourage them to be repeat customers, whereas the lost customers should be offered with promotional offers in order to re-acquire these customers.

*Central Zone:*

It is too early to analyze the central zone. But since this seems like a new set of chain stores, it is a good strategy to provide offers to all customers, but depending on items. Providing offers to selective items will encourage customers to buy particular items. Although it focuses on only particular items, it will promote the overall store sales and help in creating repeat customers.

*Root : An online Ecommerce store*

Too early to analyze. Strategies similar to the central zone have to be implemented.

## Conclusion:

In conclusion, this project involved analyzing a large dataset of a textile retail store to segment the customers into different categories based on their purchase behavior. The data was wrangled, and exploratory data analysis was carried out to understand the relationships between features. Based on insights gained, strategic recommendations were made for stores in different zones. Finally, machine learning models were built to predict customer behavior and classify them into different categories.

We utilized the XGB Classifier model as our final prediction model. XGB Classifier is an implementation of the gradient boosting decision tree algorithm, which is a machine learning technique used for classification and regression tasks. It is a popular and powerful algorithm that can handle a large amount of data and is capable of providing good accuracy.

The XGB classifier model is considered a strong model because it is an optimized implementation of gradient boosting algorithm, which is a powerful machine learning technique for predictive modeling. XGB classifier is based on an ensemble of decision trees and it is designed to be highly scalable, accurate, and fast.

The working of XGB classifier model can be summarized as follows:

1. It starts by building a decision tree with one node, which is the root node.
2. The tree is trained using a gradient descent algorithm to minimize a loss function, which measures the error between predicted and actual labels.
3. The loss function is calculated using a combination of different metrics such as mean squared error or log loss, depending on the problem at hand.
4. Once the first decision tree is built, the residuals (i.e., the differences between actual and predicted labels) are calculated and used as the target variable for the next tree in the ensemble.
5. The subsequent trees are trained using the same gradient descent algorithm, but with the added constraint that they must predict the residuals of the previous tree instead of the original labels.
6. The final prediction is obtained by summing the predictions of all the trees in the ensemble.

Our XGB Classifier model displayed an accuracy of 75% on the train set and 72% on the test set, which was the highest among the chosen models. This indicates that the model was able to accurately predict the customer category for the majority of the customers in our dataset. However, it also suggests that there may be some room for improvement in terms of the model's performance.

Overall, the XGB Classifier model proved to be a useful tool for predicting customer categories in the context of our textile retail store data. Its ability to handle large amounts of data and provide high accuracy make it a valuable addition to the machine learning toolkit for retail data analysis.

**Works Cited**

GeeksforGeeks - https://www.geeksforgeeks.org

Towards Data Science - https://towardsdatascience.com

Cats on a Hot Tin Roof: Cats Encoding Methods - https://www.kaggle.com/code/arashnic/cats-on-a-hot-tin-roof-cats-encoding-methods/notebook