Retail Sales Analysis and Customer Segment Prediction

EDA and ML Classification

Rajathadri A S

Summary of EDA

A large dataset with 1.1 Million rows and 10 features was provided to us.

Since my system could only import the dataset, but did not have the requirements to perform operations like iterations, we took a small part of 2023 dataset .

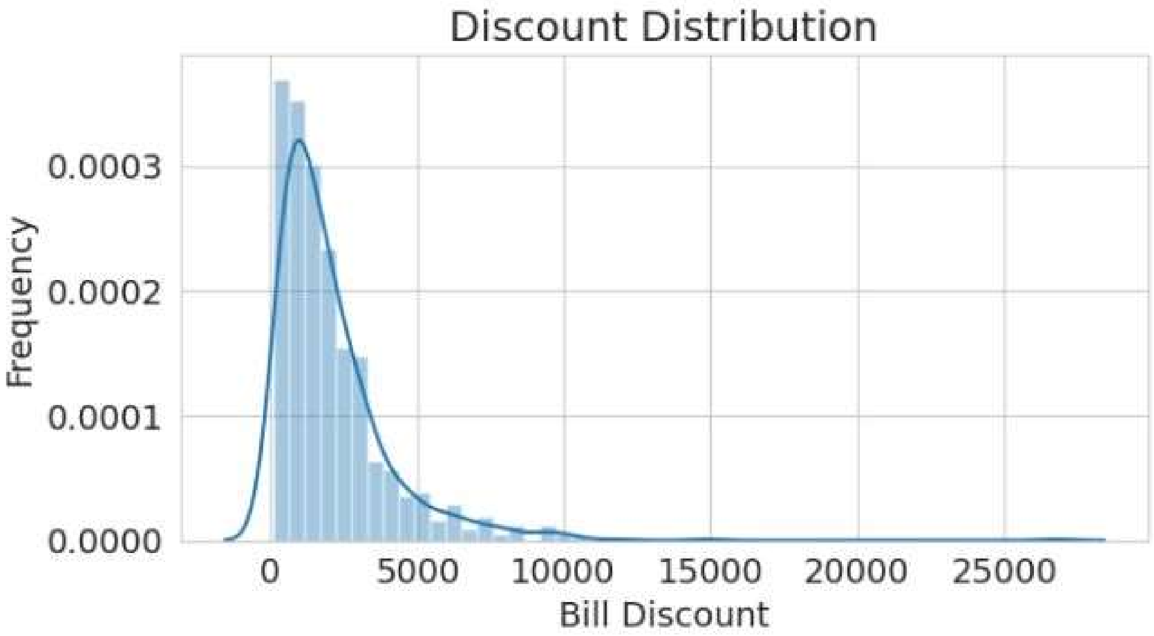
Final dataset contained 30000 rows and 10 features.

There were numerous duplicate entries and null values in our datasets, which were handled.

We performed RFM segmentation to rank customers based on their Recency, Frequency and Monetary Abilities.

We then understood our dataset using visualizations.

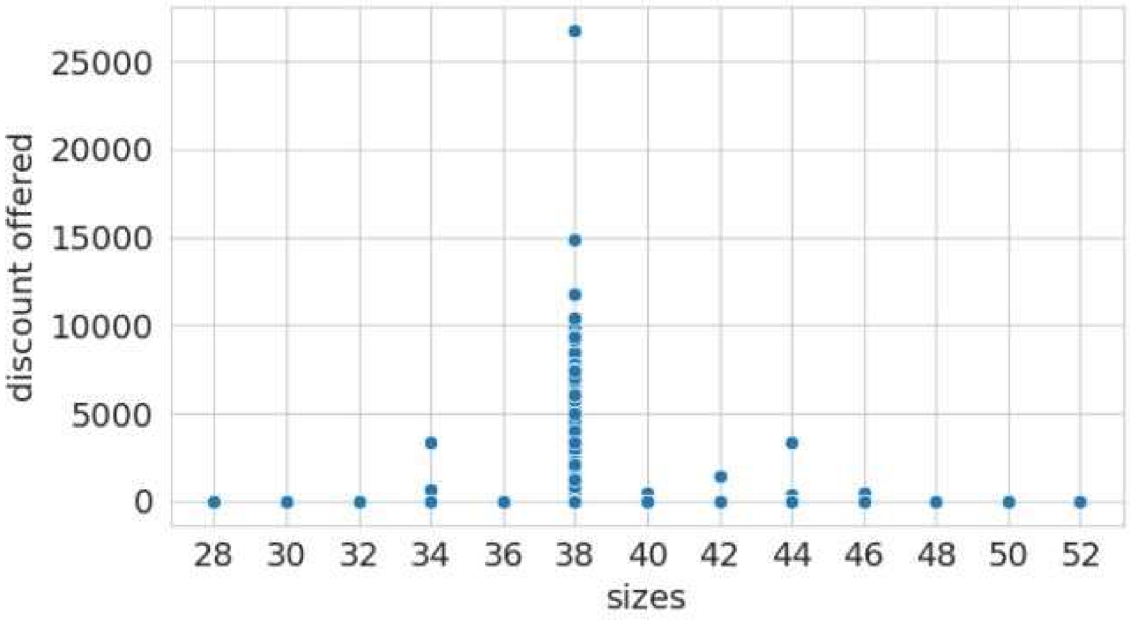
Discount



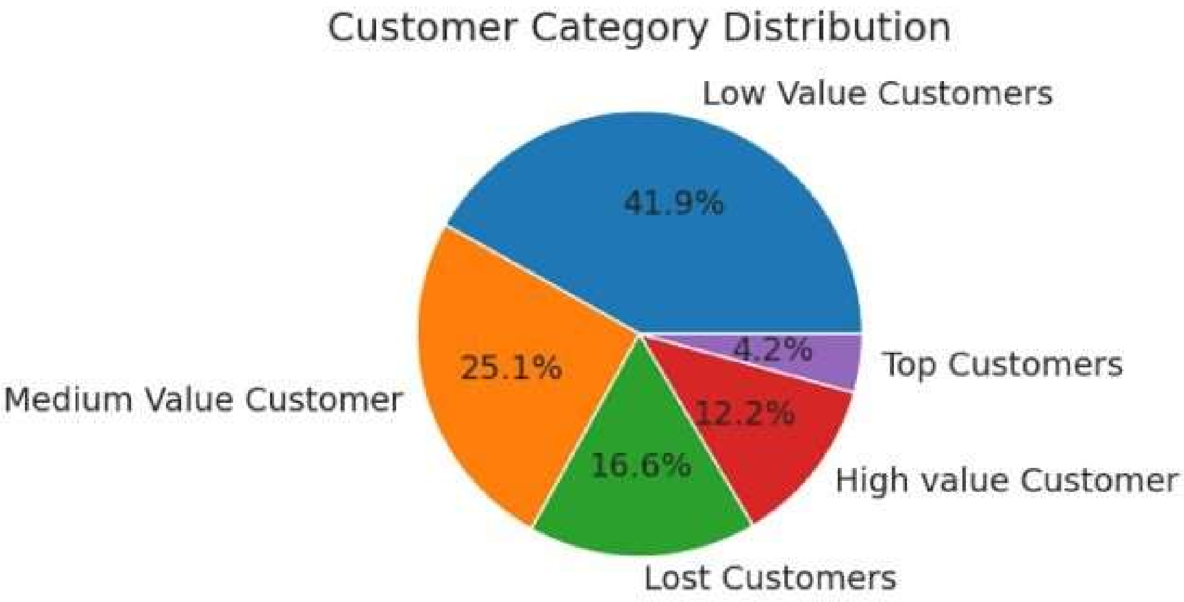
The discounts offered had a mean value of ₹2112/-

Discount vs Size

The discounts offered seem to be more focused on items of size 38, whereas other size items are barely receiving discounts.



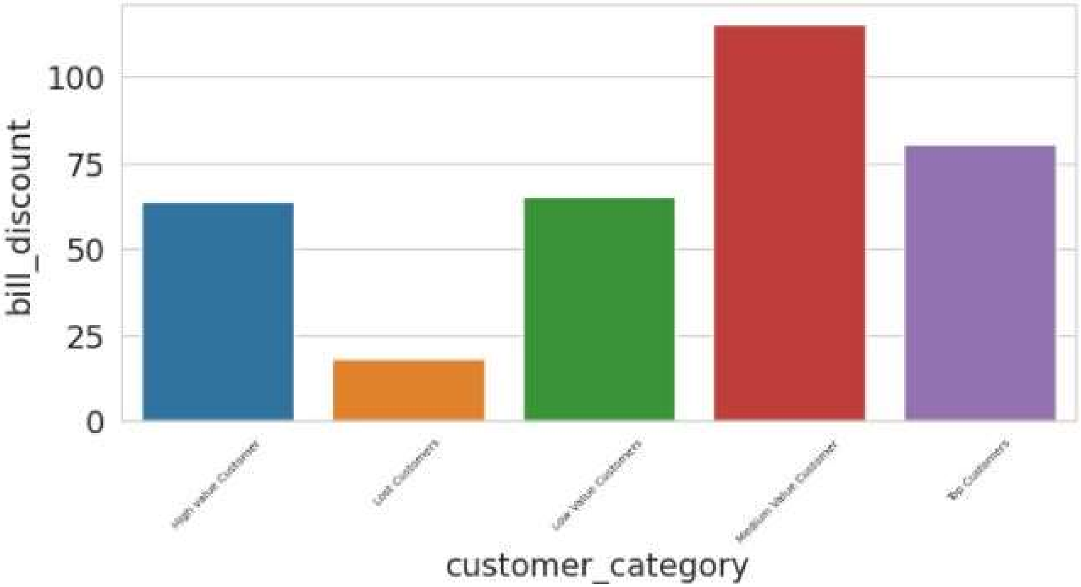
Customer Segments



From the above pie chart, we can observe that we have a majority of customers in the "Low value customers" category (about 41.9%). This is almost 2/5 of the entire customer base.At the same time, we can also observe a significant sector of customers in "Lost Customers" category as well.

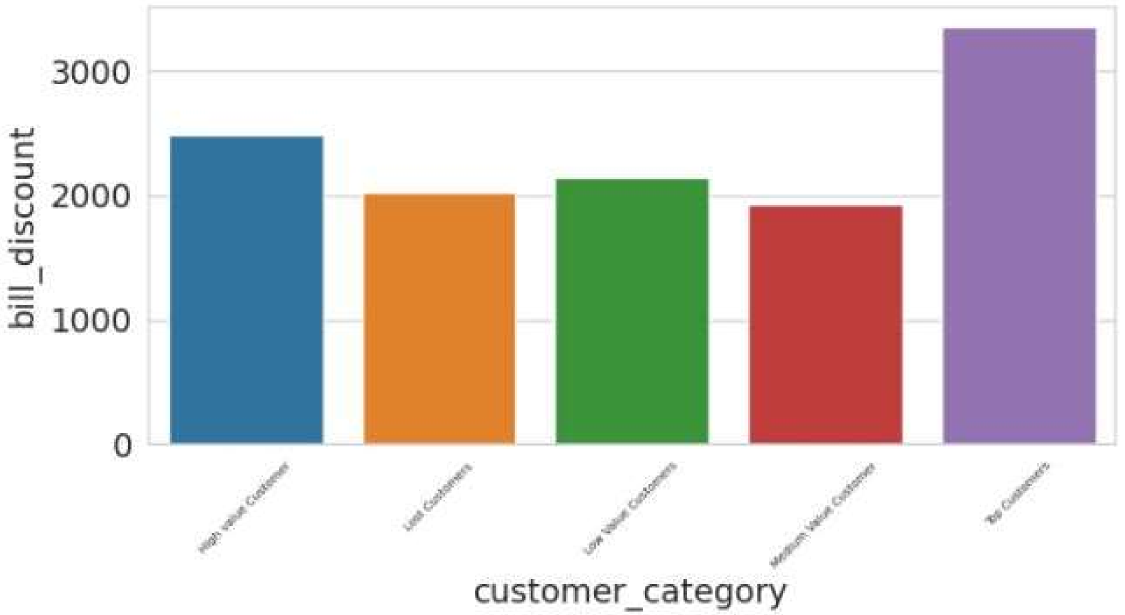
Discount per Customer segment (All customers)

We can observe that the Medium category was offered the most discount, whereas the Lost customers category was offered negligible discount when this is the category that requires the most discounts to encourage more customer retention.



Discount per Customer segment (Discounted customers only)

It is evident from the above bar plot comparison that all the customer categories have received similar discounts, except for top customers. This tells us that discounts were generally offered within a constant range to all customers, but mainly in the top-customers category.



Customer Segment Purchases per Zone

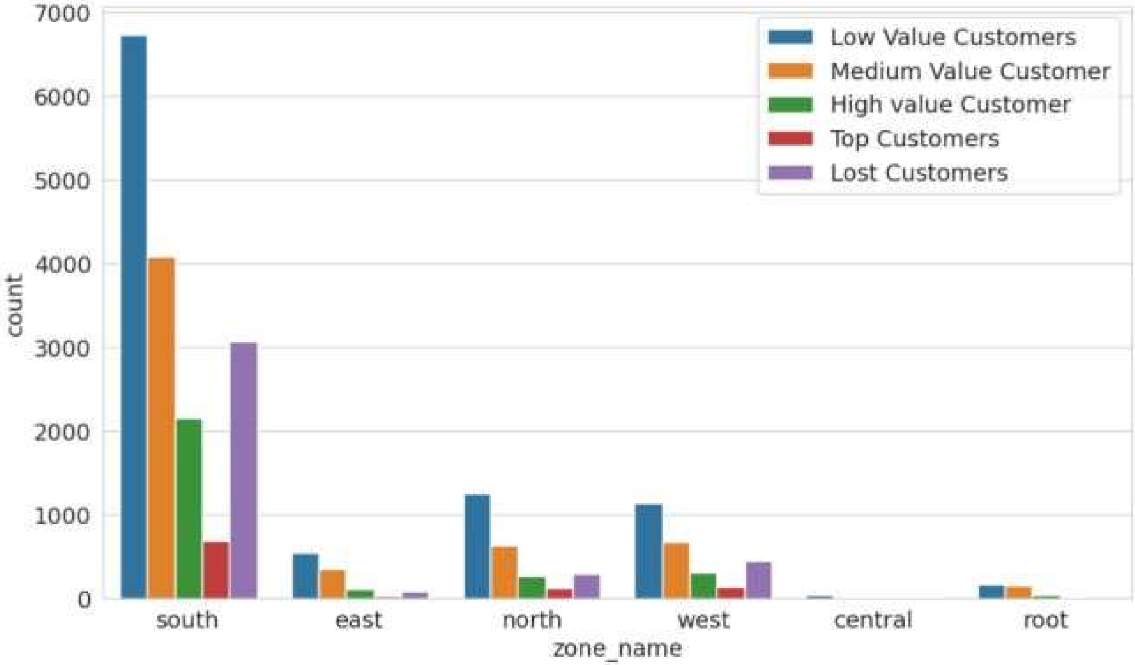
South Zone : Has the highest number of sales. These sales are mainly obtained through Low, Medium and Lost customers. It is not an ideal scenario for any sales platform in the long run.

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 customer segment.

North Zone & West Zone : These stores have a decent sales numbers, but they need to focus more on Low and Lost Customers.

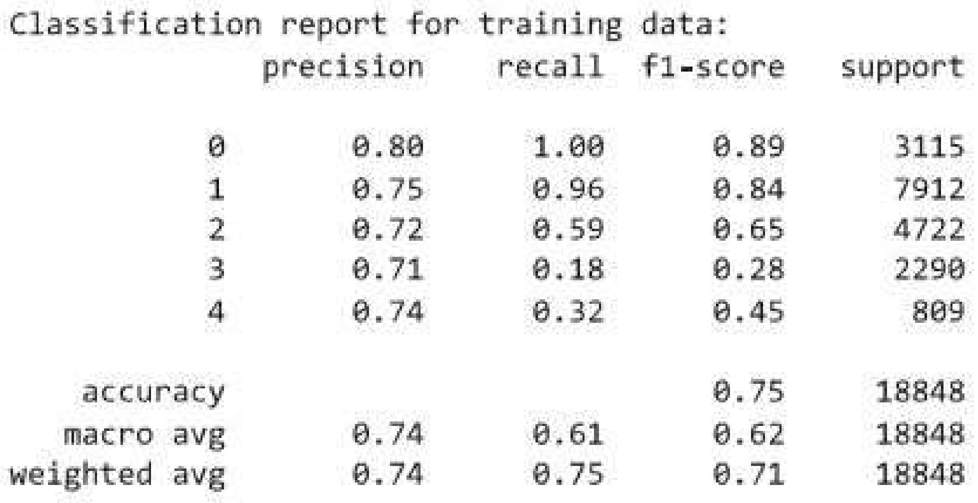
Central Zone: It is too early to analyse the central zone. But since this seems like a new set of chain stores, it is a good strategy to provide offers to all customers.

Root: An online Ecommerce store



Selected ML Classifier Model

XGB Classifier Model With RandomizedSearchCV

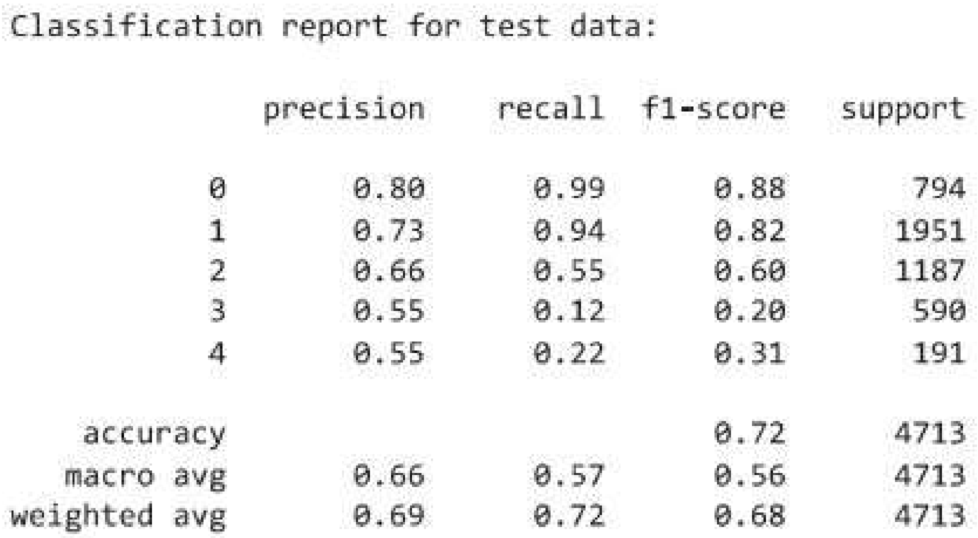


```
Classification report for training data:
              precision    recall  f1-score   support

           0       0.80      1.00      0.89      3115
           1       0.75      0.96      0.84      7912
           2       0.72      0.59      0.65      4722
           3       0.71      0.18      0.28      2290
           4       0.74      0.32      0.45       809

    accuracy                           0.75     18848
   macro avg       0.74      0.61      0.62     18848
weighted avg       0.74      0.75      0.71     18848
```

Model performance on Train set.

Selected ML Classifier Model

XGB Classifier Model With RandomizedSearchCV



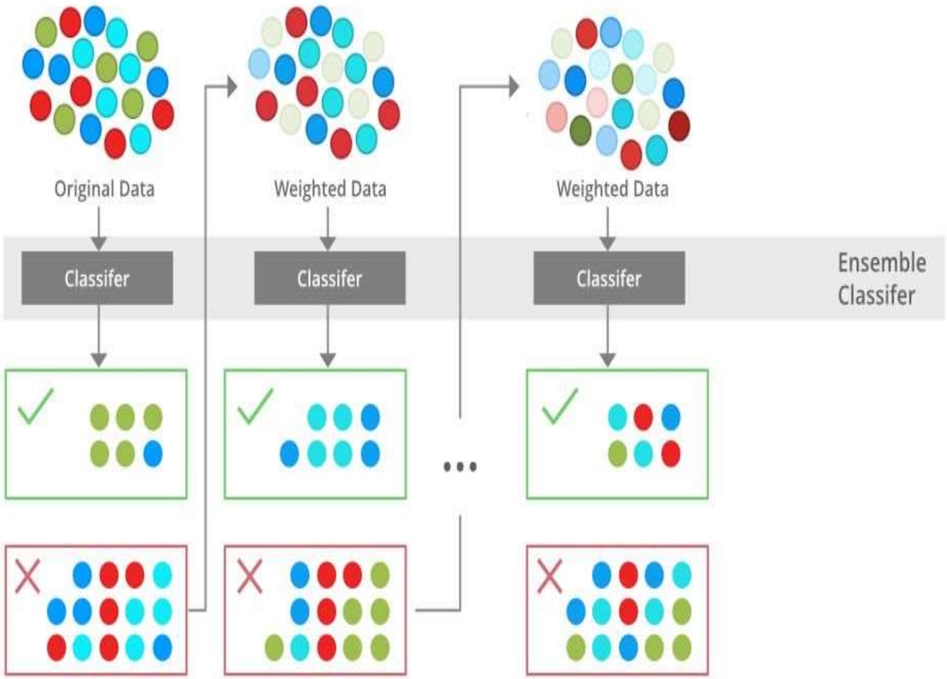
```
Classification report for test data:

              precision    recall  f1-score   support

           0       0.80      0.99      0.88       794
           1       0.73      0.94      0.82      1951
           2       0.66      0.55      0.60      1187
           3       0.55      0.12      0.20       590
           4       0.55      0.22      0.31       191

    accuracy                           0.72      4713
   macro avg       0.66      0.57      0.56      4713
weighted avg       0.69      0.72      0.68      4713
```

Model performance on Test set.

Selected ML Classifier Model

XGB Classifier Model With RandomizedSearchCV

XGB Classifier model has an accuracy of 75% on the train set and 72% on the test set. This indicates that the model is 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.



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.

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.

End of Project

Thank you.