Accusaga Assignment

Name: Muzaffar Ahmed

1. **Problem Statement:**

It is binary classification problem to predict whether customer purchased ABC product or not.

1. **Metrics**:

Metrics I have used here is

1. Accuracy
2. AUC score
3. Confusion Matrix
4. **Data Overview**

There are 30829 data points and 7 features.

There are 30829 data points and 7 Features.

There are 6 independent features and 1 Dependent Feature.

Class 0 means customer didn't purchased product.

Class 1 means customer purchased product.

It is Balanced Dataset.

Independent feature Datatype

1. Customer Id Int32
2. Var1 Float64
3. Var2 Float64
4. Group Object
5. Category Object
6. Rating Object
7. **Missing Values**

There are no Missing values and duplicates in the dataset

1. **Exploratory Data Analysis.**
2. Few outliers have been detected in Var1 and Var2 feature. Those outliers have been removed. Outliers are detected by Box plot.
3. Most of the Var1 values lies in range of 0 to 500.
4. Most of the points of Var2 lies in range of 0 to 50.
5. Mean and maximum value of Var1 and Var2 features of customers who purchased product is greater than customers who didn't purchased products.
6. Quantiles of Var1 and Var2 for customers who purchased products is greater than customers who don’t purchased.
7. 90% of customers who purchased products have Var1 less than or equal to 250 and greater than 100.
8. 95 % of customers who didn't purchase product have variance1 less than 100.
9. CDF distribution of Var1 are well separated for both classes.
10. We can say that customers who purchase ABC product will have Var1 greater than 250 and customers who didn't purchased ABC product will have Var1 less than 250.
11. 90% of customers who didn't purchased ABC product have variance2 less than or equal to 20.
12. 90% of customers who purchased ABC products have Var2 less than or equal to 40.
13. We can roughly estimate that Customer who purchased product have Var2 greater than 40 and customers who didn't purchased products have var2 less than 40.
14. Customers who didn't purchased ABC product will have low value of Var1 and Var2.For the customers who purchased product will have high value of Var1 and Var2.
15. There is strong correlation between Var1 and Var2.
16. There are 8 Categories in Group Feature. G1 and G5 are dominant categories
17. Customers who are having G3 category tends to purchase the product.
18. There are 5 Categories in Category Features. C1 and C3 are dominant categories.
19. If the customer is of category C1,C3 or C4, then customer probably will not purchase the product.
20. If the customer is of category C2 or C5, then customer tends to buy the product.
21. There are 3 Categories in Rating Feature namely Gold, Silver and Bronze.
22. Bronze and Silver Rating are dominant categories in Rating Feature.
23. If a customer is having Bronze rating, customer probably will not buy the product.
24. If a customer is having Gold or Silver rating, customer tends to buy the product.
25. **Splitting Dataset**.

Dataset is splitted in 80:20 ratio for train and test set with stratify distribution.

1. **Categorical Encoding**

One hot encoding for Category and Group Feature.Ordinal Encoding for Rating Feature.

1. **Feature Scaling**

Standard Scaling is done on both train and test dataset.

1. **Modelling**

I have trained our data on different models with Hyperparameter tuning and cross validation.

Following are the models which I have trained:

1. Logistic Regression
2. Naïve Bayes
3. Decision Tree
4. Random Forest
5. Xgboost

Out of all the models, best model is XGBOOST. Following are the results achieved from Xgboost.

1. AUC score achieved is 0.8532
2. Accuracy achieved is 75%
3. True Negative=2126, True Positive= 2515, False Negative=590, False Positive=935
4. **Feature Importance**

Feature Importance is estimated from Random Forest.

Top most Important features are

1. Var2
2. Var1
3. Rating (Ordinal Encoded)
4. C5
5. C1
6. C2
7. **Group to reach out to increase sale of product.**
8. Customers who are having Var1 value greater than 250 will have more tendency to buy product.
9. Customers who are having Var2 value greater than 35 will have more tendency to buy product.
10. Customers who are having G3 category in Group feature will have high tendency to buy product
11. Customers who are having C2 or C5 as category in Category feature, then those customers will have high tendency to buy the product.
12. If the customer is having gold or Silver rating, then customer will have high chance to buy the product