**PROBLEM 1 - Vivek**

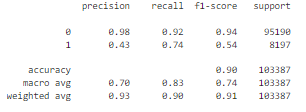
**Problem Statement:** The client is an online ecommerce player. They have shared a user level log data for their browsing behavior (Timestamp, UserId and website section visited). You have been provided with the final conversion data too (Timestamp, UserID, Products Purchased in the transaction, Overall Cart Value). Based on this data you are required to build features and model on these characteristics of users to calculate a score/rank for conversion probability of that user.

**Approach:**

Here we first summed the data to the user\_id level. We then joined the two datasets on a user\_id level. We first ran Logistic regression to predict the conversion probability, since we have to predict if a user buys a product or not. We also ran a Gradient Boost classifier to check if the results improve or not.

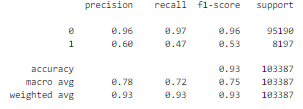
**Results:**

We found that in the case of Logistic regression:



The overall test accuracy of Logistic is 90.1%

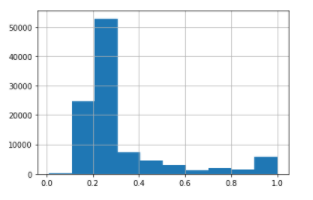
Same with GBMClassifier we get:



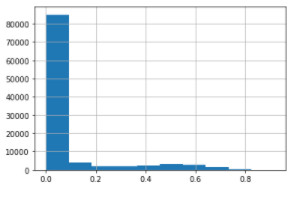
The overall test accuracy of Logistic is 93.3%

We see here that the precision for the class 1 has improved, but the recall has reduced.

The probability distribution for logistic regression is as follows:



For GBM is as follows:



Clearly we can see that Gradient boost has a more skewed distribution, i.e it has assigned a lot of values with a probability very close to 0.

Although not ideal we would go with a logistic regression as it gives results close to GBM classifier but is a much more simpler method. If we had access to more variables concerning the user’s profile, location, device, past purchase history. We also only had one day’s data to train the model.