

## Contextual Explanation:

This prediction project is done in collaboration with our client Social Insider. Social Insider provides business insights to marketing teams in large corporations by offering data analysis and comparison across business social media accounts. They have provided a 14-days free trial for trials before official subscription to their service. Currently, they would like to predict what conditions lead to the conversion of their free users in purchasing subscriptions. Our classification machine learning model targets at correctly predicting these converted subscribe users.

It was a common agreement reached during our discussion with client Social Insider that Recall is the performance matrix we should optimize for. The core of the reason lies in the formula difference between recall and precision:

$$\text{Recall} = TP / (TP + FN), \text{ and}$$
$$\text{Precision} = TP / (TP + FP),$$

Where:

TP = True Positive (Users who would have subscribed and has correctly been predicted as positive subscribe user)

FN = False Negative (Users who would have subscribed and has incorrectly been predicted as non-subscribe user)

FP = False Positive (Users who would NOT have subscribed and has incorrectly been predicted as subscribe user)

Optimizing recall tries minimizes False Negative (FN) which has a larger cost than False Positive (FP) in our scenario. Our current conversion rate is less than 1%. The cost of losing a user who would have subscribed (due to wrong prediction) is larger than the cost of extra marketing to users who would not have subscribed (no matter how much marketing they receive). Hence, we primarily use recall as the indicator matrix for choosing our best prediction model.

Visualization:

## Gradient Boosting Has the Best Recall Among All the Models

The models predict whether a given user is going to subscribe to our client Social Insider's subscription plan based on the series of user event logs.

