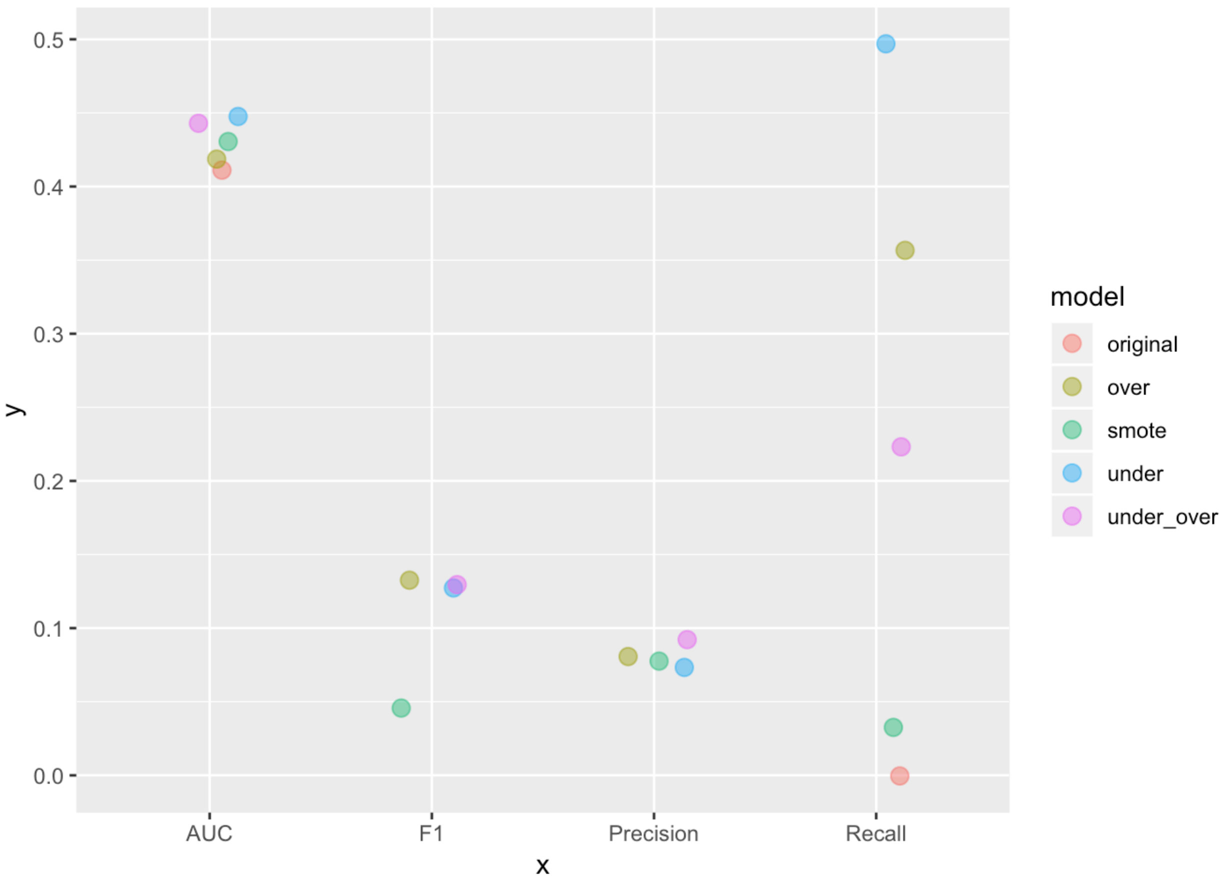
Q2 Unbalanced Dataset Classification

Model Name	Test Accuracy	Sensitivity/Recall on rare class	Test **AUC** on rare class
Baseline XGBoost	0.939	0	0.410
Undersampling	0.586	0.497	0.447
Oversampling	0.716	0.357	0.418
Under and over sampling	0.817	0.223	0.442
SMOTE	0.917	0.032	0.431

Even though Baseline model has higher test accuracy, its no better than a randomly guessing classifier since it predicted all the classes to be “OTHER”. Since our goal is to predict the “DIFFICULTY” class well, we performed imbalanced dataset classification techniques to be able to predict the minority class (‘DIFFICULTY’ target class) better. Out of the techniques used, Under sampling technique gave the highest AUC, followed by under-and-over-sampling and SMOTE.

From the plot on the next page , we can see that the recall rate – Fraction of people who have difficulty paying back loan that the classifier correctly identifies is highest for undersampling, following over sampling, followed by under-over sampling, then smote.

This shows the benefit of using these sampling methods in an imbalanced dataset like the HomeCredit Dataset.


0.5
0.4
0.3
0.2
0.1
0.0
y
AUC
F1
Precision
Recall
x
model
original
over
smote
under
under_over