**Kaggle Project 6 – Santander Value Prediction Challenge**

**Prepared by: Dwaipayan Mukherjee (2211569)**

**About the Project:**

I am provided with an anonymized dataset containing numeric feature variables, the numeric target column, and a string ID column.

The task is to predict the value of target column in the test set.

**Steps I have followed through the project:**

1. First, I have loaded the necessary libraries and the data as Pandas Dataframes.
2. I have viewed the train and test data and looked at the info about the datasets.
3. I have run a check for null values in the train and test datasets. There are no null values in either of the datasets.
4. Next, I have run a check for columns having constant values and removed them since they provide no signal for the target column.
5. Next, I ran a check for duplicate columns and deleted them from both the train and test set.
6. Next, I have removed the sparse columns from both the datasets.
7. Then, I have separated the features and target in the training data and split the train data into training and validation set.
8. Modelling:
9. LightGBM Model: I have trained the data using a LightGBM model and found the RMSE to be 1.4137.
10. XGBoost Model: I trained an XGBoost model on the data and found the RMSE to be 2.42.
11. CatBoost model: I have trained a CatBoost model on the data and found the RMSE to be 1.428.
12. Final Prediction on test set:

I have combined all the three models for my final prediction on the test set. The final prediction is a weighted average of the LGBM prediction, XGB prediction and CatBoost prediction in the proportion 0.5:0.3:0.2.