# American Express Campus Analyze This 2018

Final Submission

Credit and Fraud Risk

# Team Details

**Team Name :** **We_R_Python**

| Name | Campus | Roll No. | Mobile No. | Email Id |
|---|---|---|---|---|
| Prakhar | IIT (BHU) Varanasi | 15145033 | 7080486918 | prakhar.student.met15@itbhu.ac.in |
| Shivansh Kumar | IIT (BHU) Varanasi | 15145049 | 8601412674 | shivansh.kumar.met15@itbhu.ac.in |

# Estimation Technique Used

## Please provide the estimation/modeling technique(s)/approach used to arrive at the solution/equation

### ❑ Predicting class output:

In order to deal with the large no. of missing values, first a list wise deletion to remove the data containing large no of missing values was done, reducing the dataset to a size of 79499 rows. Then all the features containing large no. of missing values were removed, reducing the no. of columns to 35, after which all the features with high mutlicollinearity (VIF>6) were removed further reducing the no. of variables to 27. Finally after application of PCA the dataset was reduced to a (79499,19) dataset, which was trained by Light Gradient Boosting algorithm with properly tuned hyper-parameters to arrive at the solution.

### ❑ Rearrangement of Application Key:

Keeping in mid the fixed Operational Budget, rearrangement needed to be done for maximizing the final score and minimizing the operation expense. Thus in order to increase the precision and decrease the recall the threshold of probability for prediction of class 1 was set to 0.55, instead of 0.5.

# Strategy to decide final list

## Please provide the strategy employed to decide the final list for submission

- I have trained a no. of probability output classification algorithms on the training dataset. I used the following parameters to make the final choice of the model:
- 6-fold Cross validation error
- Accuracy on the test set using train-test split (I split the training data in a ratio of 70%-30% , trained the model on the 70% of the data and tested it on the left 30% of data)
- Leaderboard data score

In all of the 3 metrics, the LightGB model performed better as compared to the other models.

# Details of each Variable used in the logic/model/strategy

**Please provide details of each variable used in the final logic**

- train- training dataset
- leader- evaluation dataset
- train_1- training data after dropping features with many missing values(>20000 NAs)
- train_2 – training data after dropping highly correlated features (VIF>6)
- fin1- list of features name used for fitting the model
- train_fin- subset of original data containing only the features which are used for fitting the model
- train_fin_sub1, test_lead_1- Subset of training and evaluation dataset containing features that will be by imputed by mean impuation
- train_fin_sub2, test_lead_2- Subset of training and evaluation dataset containing features that will be by imputed by mode impuation
- train_fin_sub3, test_lead_3, test_lead_4, test_lead_5- Subset of training and evaluation dataset that contains no missing values
- train_final, test_leaderboard- Final imputed training and evaluation data
- X,y, test_1 – train data, prediction class and test data used for fitting model

# Reasons for Technique(s) Used

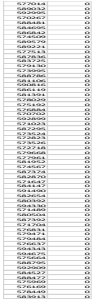
## Why do you think this is the best technique(s) for this particular problem?

I think Light Gradient Boosting was the perfect choice for this problem because:

- LightGBM is known to have a better accuracy than other algorithms due to production of much more complex trees. The only main drawback of a LightGBM is that it is prone to overfitting due to its vertical growth. Whereas in this case this drawback can easily be eliminated due to large size (around 80000 rows) and also overfitting was further controlled using proper regularization parameters.
- LightGBM is known to perform poorly if the dataset has many categorical variables. So again, in our case this limitation didn't affected the model performance due to presence of very few categorical variables.

# Final Submission File

**Please embed your final submission file (.csv) here.**



Credit and Fraud Risk

# THANK YOU


Credit and
Fraud Risk