**Overall Approach of Assignment**

**Name: Shobhit Kulshreshtha**

Environment used: Jupyter Notebook

Programming Language: Python 3

Libraries used: Numpy, Pandas, Matplotlib.pyplot, sklearn, xgboost

Detailed steps:

* Train Dataset was downloaded and studied in detail.
* Researcher analysed the features in detail and decided to use XGBoost model. It is a relatively faster Decision Tree approach and works at a faster pace. Being a Tree based model, it also does away the need to scale the variables, null value handling and outlier handling.
* At initial stage, researcher checked for duplicates that were not found in training data
* Since researcher aimed to use XGBoost, researcher skipped the null value handling and outlier handling.
* Given Train Data was split into X & y
* Features: gender, owns\_car, owns\_house were converted to bool type
* Feature occupation\_type was converted to dummy variables (drop\_first was used to drop the first dummy) and added in the X dataframe.
* X & y were split in Test and Train with 70% of data as training data.
* XGBoost.XBGClassifier was applied and initial model development was done to asses optimum epoch value to be used. Eval\_results of model were saved and plotted tp check the least difference between Test and Train and least error for test set. The plot was found to indicate that optimum results could be attained at 7 epochs. The model was hence refit with early\_stopping\_rounds at 7.
* Predictions were made on Test and Train set and difference in accuracy was found to be around 1% proving the model can be generalized.
* Test dataset was then loaded in notebook and feature engineering and transformation activities were done as done in line with our train set as indicated above.
* Null value handling, Outlier handling were not undertaken since Tree based model (XGBoost) will take care of the same.
* The model was then applied on this dataset and results were saved in a file alongwith customer\_id.

