**PROJECT: INCOME QUALIFICATION**

*Submitted by: Mr. Tejas Phase*

The whole project is divided into 2 parts.

1. **The model building without Feature engineering**.

2. **The model building with Feature engineering (dimensionality reduction using PCA)**

**How I approached the Problem:**

Without Dimensionality Reduction**:**

When done loading dataset I observed that there are 143 feature columns present in the dataset and its mix of different data types include int, float and object. Also there are some missing values present in the dataset which should be first handle before datatype conversion. So, first I identified columns which has missing values and treat them with different strategies. After dealing with all the columns with missing values I converted the categorical columns into dummy and thus all the columns were in numerical format.

After all of this conversions, I also observed that the data in the dataset was not in the Standard scale. So, I normalized the data using one of Normalization techniques. Then I separated it into features and Target and then split it into train and test dataset. Then as we prompted to use ‘Random Forrest Model’ for prediction I created its instance and trained with Training Dataset. Here, I did this step into 2 ways by changing ‘Model Parameters’ of ‘criterion’. In first step, I used default criterion which is ‘gini’ in this case and for this I evaluated the model using metrics Accuracy and then changed the criterion to ‘entropy’ and evaluated the model using metrics Accuracy. Then the same trained model was tested on Test dataset and observed the change in the Accuracy.

**Conclusion:** Due to feed all the features to model when trained it became Over-fitted and showed high variance on test dataset. To minimize this problem I used the next approach which is Dimensionality Reduction.

With Dimensionality Reduction:

As we discussed above our model was over-fitted due to provide all the features to it. So, we decided to use PCA technique of Dimensionality Reduction to reduce feature columns. So, I set no of principals equal to 52 which represented 95% of the dataset. After performed Dimensionality Reduction I got dataframe of 52 feature columns and Target column. Then I separated feature columns and Target column and performed train and test split and build model with train dataset and evaluated with accuracy metrics. Then with this trained model I performed prediction on test dataset (which was also pre-processed with PCA) and checked accuracy.

**Conclusion:** With PCA our model is so much generalized on test dataset compared with our over fitted model.

**Remark:** At this point my model needs to be retrain by doing some pre-processing on dataset. Because still there is much difference between training and testing accuracy when testing model on test dataset.