# RoboKlam AI/ML Assignment - Problem 2

By Rohit Suryavanshi
Email: rohitsurya7777@gmail.com  Phone no. : 8291942333

## Collecting and Cleaning Dataset:

- Collect Dataset on the lead score based on the call records between sales team person and customer (which will be assumed with 0% error)
- Accumulate the data within an excel or csv file, whether the entries be categorical or numerical with the lead score being binary (0 or 1).
- Use Python libraries such as numpy, pandas , seaborn, matplotlib and most important, scikit-learn to clean the dataset in a machine understandable language
    a. Convert the categorical entries to integers using One-hot encoder or create a function to do the same
    b. Perform Exploratory Data Analysis and identify the outliers that may be present within the data and remove them.
    c. Use scikit-learn module of StandardScaler() or QuantileTransformer() to normalise the training set

## Choosing the algorithm and deploying it on the dataset:

- Split the dataset into training set, cross-validation set and test set in the ratio of 60:20:20 ratio using scikit module.
- For the given problem, we are classifying whether the call is a lead or not, which is a binary classification. Hence, the algorithms to apply in such a problem are:
    1. Logistic Regression
    2. Support Vector Machines (SVMs)
    3. Random Forest Algorithm
    4. K-Nearest Algorithms

  Then we fit each of the models with our training dataset
- Compare the datasets and
- Compare the datasets accuracy using model.score(cv set) and algorithms with low accuracy can be used together using Ensemble learning methods like ensemble.Bagging.Classifier , ensemble.ExtraTrees.Classifier
- Then optimise whether to use normalisation, whether one of the cleaning techniques actually lead to downfall of accuracy etc.
- After this, we deploy the test data to calculate the accuracy of our model.

## Deep Learning - Tensorflow-keras

- Create a Neural Network Architecture using keras framework, for this classification task
- We will create multiple layers with a number of neurons depending on the number of features we would use.
- After each layer, BatchNormalization will be used and Dropout regularisation for layers with high neurons..
- Each layer will have activation function ‘ReLu’ and the model will be compiled using Adam optimizer and the losses for binary classification will be calculated using BinaryCrossEntropy.
- Now we will run the model using our cv data with epochs and batch size hyper parameterized using CV dataset and deploy the model onto our Test data.