**APPROACH DOCUMENT**

1.**Problem Statement**: “To aid staffing, we were provided with the monthly information for a segment of employees for 2016 and 2017 and tasked to predict whether a current employee will be leaving the organization.”

* The problem statement seemed like a generic classification problem statement with not a lot of variables. The task was to predict employee attrition / which employee would resign or leave the organization.

2.**Initial Approach**:

* Looking at the data set, initially there was no target variable in the training set. Reading the problem statement again, the target variable was meant to be drawn from the other variables present in the training set.

By observing the ‘**’Last Working Date**’’ variable, it was understood that the target variable was to be drawn by assigning the value ‘0’ to the blank spaces which meant that the employee hasn’t left the organization and the value ‘1’ to the spaces that were filled in with dates which meant that the employee has left on that particular date, hence creating the Target variable column.

3.**Data Preprocessing**:

1. First all the date columns such as ‘MMMM-YY’, ‘Dateofjoining’,’ LastWorkingDate’ were removed as they were not required in the model building process.
2. The column ‘**Emp\_ID**’ was only used to merge the training data and test data, to group all the columns in one variable to eliminate duplicates and to create a new data frame, after which was dropped since it was not required during the model building process.
3. The numerical columns ‘**Salary**’,’ **Total Business Value**’ and ’**Age**’ were scaled down to unit variance using the Standard Scaler function. It was a presumption that the numerical data had to be scaled since they weren’t normally distributed and vary in length from each other.
4. The categorical variables ‘**Gender’**, ‘**Education** **Level’** and ‘**City’** were Label Encoded to respective classes using the Label Encoder function as we were planning on building classification models which cannot process categorical data.
5. All NA values were dropped as it did not affect the model building
6. Up sampling was done to the Target variable using duplicate records to increase it and then concatenated into a new date frame, because of the high imbalance between 0’s and 1’s.

4.**Model Building:**

* After running the data in multiple models such as Logistic Regression, Decision Tree, Random Forest, ADA Boosting, XG Boosting, Neural Network, K Nearest Neighbor, Support Vector Machines and Naïve Bayes.
* Support Vector Machine and Random Forrest Algorithms had bought better accuracy and F1 score in comparison and was chosen for further analysis.
* Random Forest model had used 100 trees which had a maximum depth of 6 and a random state of zero to build a model with 88% accuracy and F1 score 0.66
* Support Vector Machine model had a bit of Hyper tuning such as its regularization parameter ‘C’ set to 1 and probability set to true, bringing good accuracy and higher F1 score.

**5.Conlusion:**

Support Vector Machine was considered the optimal and final model to predict the test dataset as it proved to be more accurate. Advantages of using Support Vector Machines is that the model is relatively memory efficient, that it works well and is more effective in high dimensional spaces.

The Final SVM model was reached after a lot of trial and errors using different parameters of not only SVM but other models like Random Forrest and Neural Networks.

After building multiple models and tuning different parameters, Support Vector machine model processed a higher F1 score.

**Name: DAKSHIN THOMAS TITTY GEORGE**

**Mobile Number: +91 963243360**

**Email Id: dakshintg@hotmail.com**