**PROJECT SUMMARY**

**CAPSTONE PROJECT III (CLASSIFICATION ANALYSIS) HEALTH CROSS SELL PREDICTION**

ADITYA S.REKHATE

In this project, we have to build a model that predict whether the customer is interested in taking the vehicle insurance provided by the company or not. To solve this problem we are provided with the dataset having 381109 rows and 12 columns having no null and duplicate values.

# Exploratory Data Analysis

In Exploratory data analysis, we performed univariate, bivariate the following observations were made:

* The dataset is imbalanced, having only 12.26 % of customer are interested in vehicle insurance.
* The age group between 40-50 are showing more interest in vehicle insurance than the other age groups which might for segment the customer.
* only 41 persons out of total who showed interest does not driving license.
* out of total 53 region, region having Region\_Code 28 has the maximum number of customers which is equal to 106415. And 19917 of these customers responded positively.
* Customers who have not taken vehicle insurance previously are more interested in taking the vehicle\_insurance.
* Approximately 74.5% of people who responded positively have vehicle age in between 1-2 years.
* the top 3 policy\_sales\_channel who contacted most of the customers are 152,26,54 and the top 3 through which maximum positive response were observed are 26,124,152.
* The median age of people not having driving license is approximately 65 but approximately 30 for people having driving license.
* From Customer who responded positively, The median age is approximately 55 for customer having no license and approximately 40 for having license.

# Performance metric

Performance metric used were roc\_auc,accuracy, precision, recall and confusion matrix to show the performance of model. However, for hyperparameter tuning we used roc\_auc value.

# Model training:

We trained total three models Logistic Regression, Random Forest classifier, xgbclassifier . The following observation can be made :

* The best performing model is RandomForestClassifier with hyperparameters ({'n\_estimators': 100, 'min\_samples\_split': 150, 'max\_depth': 13}). Although the train score of all the models are almost similar but test score observe was best for RandomForest classifier.