**Cap Stone Project Synopsis & the approach to detect credit card fraudulent transactions.**

**Problem Statement and the need for the project:**

In lieu of the steep increase in banking frauds , we understand that Machine learning can play a vital role in detecting fraudulent transactions and hence we do this project to use key machine learning methodologies to detect such activities.

**Key process steps that are proposed:** Below are the key process steps that can be followed for better implementation of the machine learning methodology as an approach to the problem.

* Understand the problem statement and business case
* Understand the data set – Class imbalances and way to handle
* Data Cleaning &Working on Exploratory Data Analysis – To verify the key feature variables to work with.
* Split between the Train and tested data sets – validation approach to be followed for split
* Model selection and interpretability
* Hyper parameter Tuning
* Model evaluation and performance Metrics.

Let’s get into the details step by step:

* ***Understand the problem statement and business case:*** As the banking sector is flourishing, it is getting plagued by some operational problems such as frauds etc. and hence it is very crucial to identify and eradicate such unhealthy development of fraudulent activities afflicting the banking sector which impinges their credibility adversely
* ***Understand the data set*** – **Class imbalances and way to handle:** Class Imbalance is one of the key pain points while working with huge datasets. It is very important to handle class imbalance in the current data set as over 2,00,000 cases are mapped to ‘0’, but hardly 500 are mapped to ‘1’. The loss function optimisation shall be heavily biased under this form of data and hence we need to work on techniques like Over sampling, under sampling or SMOTE which generate new points vertically between two data points that belong to such minority class. Thus, we use the K-Nearest neighbour to create random synthetic samples.
* ***Data cleaning and working on EDA:*** Finalise the variables to work with, either with time or amount of impact since the 28 variables are formed once the PCA is done.
* ***Train and Test data split:*** Splitting the data between test, train using appropriate random state values
* ***Model selection and Interpretability:*** Asit is a supervised machine learning algorithm,and a classification problem we can use the below models, but let’s understand the feasibility of using these models.

1. ***Logistic Regression:*** Model is used when the data is linearly separable and highly interpretable
2. ***Decision Trees:*** Though easily interpretable it has a major problem of overfitting
3. ***K nearest Neighbour:*** though this might give accurate results but its time consuming as we need to calculate distance from all data points
4. ***Random forests:*** As the data is structured for better accuracy, we use XG Boosting methodology for number of trees using classifier as it is high-performing models.

* ***Hyper parameter tuning:*** This is proposed to be done as it improves the model performance to the next level causing cost savings. As the number or fraudulent transactions are less, we shall use stratified K Fold cross validation along with Grid search CV as the class imbalance is high.
* ***Model evaluation and performance metrics:*** Though the model might get an accuracy percentage of 99.83%, but this might not be the real situation as the classes are highly imbalanced and hence accuracy may not be the right method to evaluate the model. Thus, we need to look into other metrics like Precision, Recall and AUC-ROC score for evaluating the model.