[Project](https://www.dezyre.com/project/hackerday-project/project-title/human-activity-recognition) Overview

**Business Context** - Banks are primarily known for money lending business. The more money they lend to people whom they can get good interest with timely repayment, the more revenue is for the banks. This not only save banks’ money from having bad loans but also improves image in the public figure and among the regulatory bodies.

The better the banks can identify people who are likely to miss their repayment charges, the more in advance they can take purposeful actions whether to remind them in person or take some strict action to avoid delinquency.

In cases where a borrower is not paying monthly charges when credit is issued against some monetary thing, two terms are frequently used which are delinquent and default.

**Delinquent** in general is a slightly mild term where a borrower is not repaying charges and is behind by certain months whereas **Default** is a term where a borrower has not been able to pay charges and is behind for a long period of months and is unlikely to repay the charges.

* This case study is about identifying the borrowers who are likely to default in the next two years with serious delinquency of having delinquent more than 3 months.
* We have a general profile about the borrower such as age, Monthly Income, Dependents and the historical data such as what is the Debt Ratio, what ratio of amount is owed with respect to credit limit, and the no of times defaulted in the past one, two, three months.
* We will be using all these features to predict whether the borrower is likely to default in the next 2 years or not having delinquency of more than 3 months.
* These kinds of predictions will help banks to take necessary actions.
* **Objective** : Building a model using the inputs/attributes which are general profile and historical records of a borrower to predict whether one is likely to have serious delinquency in the next 2 years

We will be using Python as a tool to perform all kind of operations.

Main Libraries used –

* **Pandas** for data manipulation, aggregation
* **Matplotlib** and **Seaborn** visualization and behavior with respect to the target variable
* **NumPy** for computationally efficient operations.
* **Scikit Learn** for model training, model optimization and metrics calculation.
* **Imblearn** for tackling class imbalance problem.
* **Shap** and **LIME** for model interpretability
* **Keras** for Neural Network(Deep Learning architecture)

**What will you learn** – 

* Understanding the problem statement
* Understand the dataset and the behavior of the features/attributes.
* Performing Exploratory Data Analysis to understand how the data is distributed and what is the behavior of the inputs with respect to target variable which is SeriousDelinquencyin2Years.
* Data preprocessing will be one based on how the values are distributed such as are there any data entry errors that needed to be removed, outlier treatment, which is necessary for certain algorithms, imputing missing values if there are any.
* Splitting dataset into the train and test dataset using Stratified Sampling to maintain the event rate across the different datasets so that a model can learn behavior from the training dataset and can predict with certain accuracy up to some on the unseen dataset.
* Feature Engineering for better decision making by a model.
* Scaling of the features using BoxCox transformation and Standardization.
* Training a model using Neural Network as a Deep Learning architecture and analyzing the impact of training on same dataset yet having different features input values because of scaling features, increasing and decreasing minority class.
* Training a model using statistical technique Logistic Regression and analyzing why scaling features is necessary in such statistical techniques.
* Training a model using Tree based algorithms such as Bagging and Boosting and analyzing why certain techniques are not required for such algorithms which are quintessential in other modeling techniques.
* Hyperparameter tuning of the modeling algorithms and checking its impact on model performance
* Using Recursive Feature Elimination using Cross Validation to check whether any highly correlated features are there in the model and what are the optimal no of features to be used for training.
* Analyzing why a popular metric Accuracy will not be useful in our case
* Checking the model performance on the unseen dataset using metrics such as F1 score, Precision, Recall and the AUCROC
* Model Interpretability using SHAP at a global level and LIME at a local level.