Any commercial bank will receive multiple requests for credit cards from various peoples. These people may have different background with respect to their finances. They may have high loan balances, low salary etc. Ideally companies would like to give credit card to those people who would pay bill on time. Manually checking the application for credit card might be tough considering the volume of application. However, the same can be achieved using concepts of Data Science and applying a few algorithms in the dataset.

In this project, we are going to do something similar. Please follow the instructions below to carry out this project.

Note: Please note that, you have to understand each and every step before you move to next step. Whatever you are doing, you should know why you are doing and then how to do it? All the concepts should be crystal clear – then only projects like these would be of your help.

1. Load the dataset
2. Extract summary statistics of the data
3. Inspect the missing values in the dataset
    a. Replace questions marks with NaN
4. Impute NaN with mean – you are learning imputation here. Explore other methods of imputation. What kind of imputation would be used in what situation?
   An important question that gets raised here is *why are we giving so much importance to missing values*? Can't they be just ignored? Ignoring missing values can affect the performance of a machine learning model heavily. While ignoring the missing values machine learning model may miss out on information about the dataset that may be useful for its training. Then, there are many models which cannot handle missing values implicitly such as LDA.
5. Imputation might not work for all columns. Check for which columns it's not working. We need to do something else to deal this prob.
   We are going to impute these missing values with the most frequent values as present in the respective columns. This is good practice when it comes to imputing missing values for categorical data in general.
6. Pre-processing the data
    a. The missing values are now successfully handled.

       There is still some minor but essential data preprocessing needed before we proceed towards building our machine learning model. We are going to divide these remaining preprocessing steps into three main tasks:

        1. Convert the non-numeric data into numeric.
        2. Split the data into train and test sets.
        3. Scale the feature values to a uniform range.

       First, we will be converting all the non-numeric values into numeric ones. We do this because not only it results in a faster computation but also many machine learning models (like XGBoost) (and especially the ones developed using scikit-learn) require the data to be in a strictly numeric format. We will do this by using a technique called Label encoding.

7. Splitting the dataset into train and test sets
   We have successfully converted all the non-numeric values to numeric ones.
   Now, we will split our data into train set and test set to prepare our data for two different phases of machine learning modeling: training and testing. Ideally, no information from the test data should be used to scale the training data or should be used to direct the training process of a machine learning model. Hence, we first split the data and then apply the scaling.
   Also, features like DriversLicense and ZipCode are not as important as the other features in the dataset for predicting credit card approvals. We should drop them to design our machine learning

model with the best set of features. In Data Science literature, this is often referred to as *feature selection*.

8. Pre-processing the data
   The data is now split into two separate sets - train and test sets respectively. We are only left with one final preprocessing step of scaling before we can fit a machine learning model to the data.
   Now, let's try to understand what these scaled values mean in the real world. Let's use CreditScore as an example. The credit score of a person is their creditworthiness based on their credit history. The higher this number, the more financially trustworthy a person is considered to be. So, a CreditScore of 1 is the highest since we're rescaling all the values to the range of 0-1.
9. Fitting a logistic regression model to the train set – concept of logistic and other regression and classification technique should be clear
10. Making predictions and evaluating performance
    Evaluate the model on the test set with respect to classification accuracy. Also take a look the model's confusion matrix.
    In the case of predicting credit card applications, it is equally important to see if the machine learning model is able to predict the approval status of the applications as denied that originally got denied. If the model is not performing well in this aspect, then it might end up approving the application that should have been approved. The confusion matrix helps us to view our model's performance from these aspects.
11. Grid Searching and making the model perform better
    What's the accuracy in step 10?
    Now do a grid search of the model parameters to improve the model's ability to predict credit card approvals.
    You can do grid search over:
    - tol
    - max_iter
12. Find the best performing model
    Once the grid of hyperparameter values is defined and converted them into a single dictionary format which GridSearchCV() expects as one of its parameters. Now, begin the grid search to see which values perform best.
    Instantiate GridSearchCV() with earlier logreg model with all the data.
    Instead of passing train and test sets separately, supply X (scaled version) and y.

    perform a cross-validation of five folds using GridSearchCV

    What's the best-achieved score and the respective best parameters?

    While building this credit card predictor, we tackled some of the most widely-known preprocessing steps such as **scaling**, **label encoding**, and **missing value imputation**. We finished with some **machine learning** to predict if a person's application for a credit card would get approved or not given some information about that person.
    A lot of concepts are involved over here, you will have to do a lot of research before applying to this. Please use internet extensively to carry out each step and understand the underlying concept behind those. You may like to change your approach apart from what is mentioned in the steps – but it should have a logic behind it.