# Capstone Project Proposal

*Machine Learning Engineer Nanodegree*

## 1. Domain Background

Financial problem is one the serious problems that may drag people from heaven back to earth. For example, some people may quit college because of the tuition or supporting the family. We may realize that a term deposit which is like a self-funded insurance and a back-up plan may help avoid or reduce the influence of them.

## 2. Problem Statement

We would like to build classifiers based on supervised learning algorithms to make classification about if a client has subscribed a term deposit.

## 3. Datasets and Inputs

The dataset is obtained from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing) and contains 45211 observations and 17 variables.

- age: how old the client is?
- balance: the current balance in the bank
- duration: last contact duration, in seconds (numeric). **Important note**: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
- campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric)
- previous: number of contacts performed before this campaign and for this client (numeric)
- job: type of job ("admin.", "blue-collar", "entrepreneur", "housemaid", "management", "retired", "selfemployed", "services", "student", "technician", "unemployed", "unknown")
- marital: marital status ("divorced", "married", "single", "unknown"; note: "divorced" means divorced or widowed)
- education: education level ("secondary", "tertiary", "primary", "unknown")
- default: has credit in default? ("no", "yes")
- housing: has housing loan? ("no", "yes")
- loan: has personal loan? ("no", "yes")
- contact: contact communication type ("cellular", "telephone", "unknown")
- poutcome: outcome of the previous marketing campaign ("failure", "other", "success", "unknown")
- **RESPONSE** - has the client subscribed a term deposit? (binary: "yes", "no")

## 4. Solution Statement

We would like to build classifiers based on supervised learning algorithms to make classification about if a client has subscribed a term deposit. For supervised learning, I would like to perform `Logistic Regression Classifier`, `K Nearest Neighbors Classifier`, `Random Forest Classifier`, etc. After this, I also would like to use `Multilayer Perceptron Neural Network` and check its performance.

## 5. Benchmark Model

Ususally, older people tend to have a higher willing to subscribed term deposit since they are less open to the risk because of family, physical issues and/or some other reasons. Hence, our benchmark model would be set based on variable age only. If age is greater than 65, we predict those people have subscribed term deposit already. If age is less than or equal to 65, we don't think they have subscribed it yet.

## 6. Evaluation Metri

`Accuracy`, `precision` and `recall` are the metrics that we are about to use here. All these three metrics range from 0 to 1. The higher the values is, the better our model is. We can read that `accuracy` mainly focuses on the overall correct classification for both positive and negative label while `precision` and `recall` concentrate on correct classification for positive label (in this project it means `y` equal to "yes") only.

## 7. Project Design

The Project will go through these parts:

- Introduction: Briefly introducing the background, datasets and other related information
- Data Preprocessing: Detailingly analysing the dataset, including missing value/outlier detection, data visualization, feature selection, etc.
- Model Fitting: Building models based on several machine learning algorithms, including but not limited to Logistic Regression Classifier, Support Vector Classifier, Decision Tree Classifier, etc.
- Comparison: Making predictions and corresponding accuracies and AUCs of models that we fit and find the one with the most satisfactory performance.
- Summary and future thoughts
- Reference