**TEAM PROPOSAL**

TEAM-11: 1. Anjali Mudgal

2. Medhasweta Sen

3. Guoshan Yu

Our goal as a team is to create a project where we can obtain, pre-process, and clean a dataset we found online and present an analysis that includes EDA, and various model building tasks that apply to the dataset.

Modern technology has made customers aware of the developments in the economic environment, which includes the financial system. The financial needs of the customers have grown multifold into various forms like quick cash accessibility, money transfer, asset security, increased return on surplus funds, financial advice, deferred payments, etc. With a wide network of branches, even in a dissimilar banking scenario, customers expect the banks to offer a more and better service to match their demands and this has compelled banks to take up marketing in right earnest. With a wide network of branches, even in a dissimilar banking scenario, customers expect the banks to offer a more and better service to match their demands and this has compelled banks to take up marketing in right earnest. Marketing is the process of introducing and promoting a product or service to the market and encouraging sales from the buying public. Good marketing has become an increasingly vital ingredient for business success. Hence, the data set we have chosen is related to direct marketing of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact with the same client was required, to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The classification goal is to predict if the client will subscribe to a term deposit (variable y).

We have obtained our dataset from UCI Machine Learning Repository. The link for which has been attached here with: <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#>

The dataset has 45.2k observations and 21 variables. They are as follows:

1 - age (numeric)  
2 - job : type of job (categorical: 'admin.' ,'blue collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')  
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)  
4 - education (categorical:'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown)  
5 - default: has credit in default? (Categorical: 'no','yes','unknown')  
6 - housing: has housing loan? (Categorical: 'no','yes','unknown')  
7 - loan: has personal loan? (categorical: 'no','yes','unknown')  
# related with the last contact of the current campaign:  
8 - contact: contact communication type (categorical: 'cellular’, ‘telephone')  
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')  
10 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')  
11 - 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.  
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)  
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)  
14 - previous: number of contacts performed before this campaign and for this client (numeric)  
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')  
# social and economic context attributes  
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)  
17 - cons.price.idx: consumer price index - monthly indicator (numeric)  
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)  
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)  
20 - nr.employed: number of employees - quarterly indicator (numeric)

21 - y - has the client subscribed a term deposit? (Binary: 'yes’, ‘no')

SMART questions:

1.Relationship between subscribing the term deposit and how much the customer is contacted (last contact, Campaign, Pdays, Previous Number of contacts)

2.Since the dataset is imbalanced, will down sampling/up sampling or other techniques improve upon the accuracy of models.

3.Marital status, age, job, and loan to find out the financially stable population? Will that affect the outcome?

4.Effect of dimensionality reduction on accuracy of the model.

5.The optimal cut off value for classification of our imbalance dataset.

6.Modeling to estimate the potential population who would subscribe to term deposit.

7. How are the likelihood of subscriptions affected by social and economic factors?

The link to our GitHub repository is: <https://github.com/yuguoshan/22FA_6103Proj_Team-11>

The modelling techniques as proposed are:

1. Logistic Regression
2. Random Forest Classifier
3. Decision Tree Classifier
4. K Nearest Neighbor Classifier