TOPICS FOR CAPSTONE PROJECT

TOPIC 1: Objective: To predict the success of telemarketing calls for selling bank long-term deposits i.e. the classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Source: The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. It is a real data collected from May 2008 to June 2013. Total observation is 52944 contacts. Each record includes dependent variable is the contact outcome – success/failure. The independent variables are as follows:

# Bank client data:

* + - 1. age (numeric)
      2. job : type of job (categorical: 'admin.', 'blue- collar', 'entrepreneur', 'housemaid', 'management',
      3. 'Retired','self-employed','services','student','technician','unemployed','unknown')
      4. marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
      5. education (categorical: basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course',
      6. 'university degree', 'unknown')
      7. default: has credit in default? (categorical: 'no', 'yes' ,'unknown')
      8. housing: has housing loan? (categorical: 'no' ,'yes' ,'unknown')
      9. loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

# related to the last contact:

* + - 1. contact: contact communication type (categorical: 'cellular', 'telephone')
      2. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
      3. day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
      4. duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no').

# other attributes:

* + - 1. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
      2. 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)
      3. previous: number of contacts performed before this campaign and for this client (numeric)
      4. poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

# Social and economic context attributes

* + - 1. emp.var.rate: employment variation rate - quarterly indicator (numeric)
      2. cons.price.idx: consumer price index - monthly indicator (numeric)
      3. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
      4. euribor3m: euribor 3 month rate - daily indicator (numeric)
      5. nr.employed: number of employees - quarterly indicator (numeric)

TOPIC 2: Home Mortgage Disclosure Act (HMDA) collects data about mortgages. It requires financial institutions to maintain, report, and publicly disclose information about mortgages.

The various usefulness of this data are:

1. Helps lenders serve community needs.
2. Informs policy and decision making.
3. Shed light on lending patterns. Helps identify possible discriminatory lending practices.

The data includes various features such as:

1. It includes information related to loan: Mortgage applications irrespective of whether it is accepted or rejected; Loan amount; type of loan; loan purpose and reasons for denial.
2. It includes demographic information such as applicant’s race, ethnicity and sex.
3. It includes information related to lender- its name and regulator
4. Information about the property- type of property, owner occupancy, census tract.

What are the interesting questions that can be asked related to this database?

1. What are the factors that determine whether the mortgage for home purchases loans are accepted or denied? Is there any discrimination based on ethnicity, income or sex?
2. What are the factors that determine whether the mortgage for refinancing home loans are accepted or denied? Is there any discrimination based on ethnicity, income or sex?
3. What is the type of property that the application were accepted or denied based on the total application. This will give an indication of the type of property in demand.
4. What is the distribution of loan size and how it varies by property type and location? It will give us an indication about the supply of loan by amount and property type.
5. What is the distribution of lenders and the respective regulatory agencies?

Topic 3: To predict whether a person makes over 50K a year or not based on individual characteristics of profession, age, education, sex and other demographic variables. The data is based on 1994 census database. The data is extracted by Barry Becker.

Topic 4: To predict who will be interested in buying a specific insurance product and to explain why people would buy. This dataset is owned and supplied by the Dutch datamining company Sentient Machine Research, and is based on real world business data.

Each record consists of 86 attributes, containing sociodemographic data and product ownership. The sociodemographic data is derived from zip codes. All customers living in areas with the same zip code have the same sociodemographic attributes