**Homework 1**

Answer questions 1 through 5 in the Exercises section from Chapter 1 of the Larose textbook.

In your submission, you must:

* Clearly answer each question in a Word document and submit the Word document to the Homework 1 assignment.
* Write your answers completely and elaborate on your ideas as much as possible.

**1.** Refer to the Bank of America example early in the chapter. Which data mining task or tasks are implied in identifying “the type of marketing approach for a particular customer, based on the customer’s individual profile”? Which tasks are not explicitly relevant?

Ans. Following are the Data mining Task or tasks that are implied in identifying the type of marketing approach for a particular customer

* **Description** – Most of BoA customers were receiving calls from There is a need of promoting specific category of product to specific customers instead of making they need to figure out patterns and trends among their customers, instead of making 13 million calls they reduced it to specific clustered group
* **Classification -**  BoA data set was used to classify the major grouping of the customers based on their history but based on their specific “individual” interests
* **Clustering** – Clustering was mostly to group customers having similar interests and inclination towards their products
* **Association** – Only those new services of choice were offered to the customers

Here following tasks are not explicitly relevant –  
**- Prediction and Estimation** – Bank of America did not Predicted any numbers of measures which they were likely to use in the future, also they did not estimate on anything for customer choices but rather classify and cluster the group of customers for their specific product choices.

**2.** For each of the following, identify the relevant data mining task(s):

1. **The Boston Celtics would like to approximate how many points their next opponent will score against them.** 
   * **Description**
   * **Estimation**
   * **Prediction**
2. **A military intelligence officer is interested in learning about the respective proportions of Sunnis and Shias in a particular strategic region.** 
   * **Description**
   * **Estimation**
   * **Classification**
   * **Association**
3. **A NORAD defense computer must decide immediately whether a blip on the radar is a flock of geese or an incoming nuclear missile.** 
   * **Description**
   * **Estimation**
   * **Prediction**
   * **Classification**
4. **A political strategist is seeking the best groups to canvass for donations in a particular county.** 
   * **Description**
   * **Estimation**
   * **Prediction**
   * **Clustering**
   * **Association**
5. **A homeland security official would like to determine whether a certain sequence of financial and residence moves implies a tendency to terrorist acts.** 
   * **Description**
   * **Prediction**
   * **Clustering**
   * **Classification**
6. **A Wall Street analyst has been asked to find out the expected change in stock price for a set of companies with similar price/earnings ratios.** 
   * **Description**
   * **Estimation**
   * **Clustering**
   * **Association**

**3.** For each of the following meetings, explain which phase in the CRISP–DM process is represented:

1. **Managers want to know by next week whether deployment will take place. Therefore, analysts meet to discuss how useful and accurate their model is.**

**Ans. Evaluation Phase**

* After model is deployed there is a need to evaluate efficiency and effectiveness of the model.
* Hence analysts meet to discuss about the usefulness and accuracy of their model in Evaluation phase

1. **The data mining project manager meets with the data warehousing manager to discuss how the data will be collected.**

**Ans. Data Understanding Phase**

* This is initial phase where the Business problem statement is defined but there is no data related to the problem evaluation and solution in order to take forward so the project manager need to talk to the data warehousing manager to discuss and understand the basic data structure in order to discuss and identify how the data will be collected

1. **The data mining consultant meets with the vice president for marketing, who says that he would like to move forward with customer relationship management.**

**Ans. Deployment Phase**

* When the Data mining consultant meets with the vice president for marketing and when he expresses to move forward with the customer deployment of the mode that means the stage is Deployment stage

1. **The data mining project manager meets with the production line supervisor to discuss**

**implementation of changes and improvements.**

**Ans. Evaluation Phase**

* Since there are communication regarding changes and improvement of the model, there is a Evalution of model that is conducted in order to improve the model efficiency and performance effectiveness

1. **The analysts meet to discuss whether the neural network or decision tree models should be applied. \**

**Ans. Business Understanding Phase**

* This is initial phase where the team members are trying to understand the problem to define their requirements and identify which will be best suited and applicable in their case

1. **Discuss the need for human direction of data mining. Describe the possible consequences of relying on completely automatic data analysis tools.**

**Ans. Need for human direction of Data mining**

1. **Automatic Data mining algorithms not always work there is a need of human intervention, data mining is a process that improves over time with keen observation.**
2. **Every phase requires human intervention to monitor maintain and making updates to the deployed model**
3. **There is continuous evaluation of model correctness and quality measures which defines the effectiveness of any model in use.**
4. **Every model runs differently and has different benefits which can be analysed only by human observation**
5. **Combination of subject matter knowledge with specific domain matters which can only be understood and applied by humans not on autonomous machines**

**Possible consequences of relying on completely automatic data analysis tools**

1. **Automatic Data analysis tools usually refer to the older data sets which is outdated for data which is not actually used or not examined**
2. **There is no guarantee of positive results**
3. **It does not always identify the causes of business problems,**
4. **Data mining return rates and quite different and vary it does not usually repays for itself as predicted.**
5. **CRISP–DM is not the only standard process for data mining. Research an alternative methodology. (*Hint:* SEMMA, from the SAS Institute.) Discuss the similarities and dif- ferences with CRISP–DM.**

SEMMA is a sequential known as Sample, Explore, Modify , Model and Assess has similar structure to that of CRISP Phases of Data analytics in following ways –

* Initial Phase –
  + Business Understanding Phase and Sample phase are respective initial phase which helps to understand the business application of analytics find patterns, identify and define a problem statement to solve.
  + The dataset subset is selected that will be a representative of entire dataset and should contain all the information to retrieve, this helps in building a better understanding of the problem statement based on the variables and dependent variable factors for the business problem statement.
  + For Data understanding phase which falls in same lines with the Sample Phase of SEMMA process
* Exploratory Phase
  + For both the business methodologies there is an exploratory phase with title Explore and Data interpretation phase from SEMMA and CRISP respectively.
  + data understanding and interpretation phase
* Modelling, Modify, Assess phase (SEMMA)--- Modelling and Evaluation phase(CRISP)
  + CRISP and SEMMA have almost similar next stages for modelling and evaluating their respective models for their efficiency and effectiveness of use.
* Deployment and assessment
  + CRISP has a dedicated deployment phase which directly takes up the model to implementation in real world based on the statistics collected as part of model building and evaluation and testing.
  + Whereas there is no SEMMA final steps which identifies and take it officially in use, instead it is on the run implemented in the network even before testing and evaluation of the model

Difference with CRISP- DM—

* SEMMA model directly starts with the Data set and creating the respective Data set which defines the subset and identifies the respective data set with precise requirements.
* The CRISP Model instead identifies the patterns and trends to define the problem statement and then it evaluates the datasets to explore with understanding and interpretation phase
* CRISP has deployment phase which gives a good decision making strategy to identify and define a better model to deploy by pre-testing the model and confirming on its efficiency and effectivity.