**6. You will be required to present your project in class focusing on the following three items**

**1) Data description**

**• Describe your data such as where do you get it, how many records are, how many and what variables are and so on**

- Telemarketing to sell long-term deposits from Portuguese retail bank(소액 거래 은행)

- https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

- Record의 경우 총 41178개

- Variable의 경우는 임의로 추가한 Key를 포함해서 22개(Key 1개, Input 20개, Target 1개)

- Attribute의 경우 bank client data, related with the last contact of the current campaign, social and economic context attributes 이렇게 3가지로 구성이 됨(Client information, telemarketing attributes, Social and economic indicators)

- 2008년 5월 ~ 2010년 11월

**2) Classification task: Using at least three algorithms (e.g., NB, ANN, and DT)**

**• State an interesting classification problem (i.e., what you want to predict or classify)**

**O What is class variable and what is the distribution of each class**

- Yes의 경우가 Class variable

- Whether this customer will positively respond to the marketing campaign or not?

**O What is the minimum acceptable performance indicator (e.g., accuracy) from Business Intelligence algorithms**

- Prior probability가 88.73%가 나옴

**• Present what Business Intelligence models are adopted and implemented**

- NB, DT, ANN 모두 사용, Clustering도 사용함

**• Present which Business Intelligence model is the best in terms of accuracy**

DT가 가장 좋은 결과가 나옴(91.28%)

- 그런데 과연 Prior와의 3%의 차이가 실제로 유의미한가(교수님께서 물어보신다)

**• Present which Business Intelligence model is the worst in terms of accuracy**

NB가 가장 결과가 좋지 않음(86.55%) -> 이는 minimum acceptable performance indicator보다 더 낮은 결과

**• Explain why do you think a Business Intelligence model works well and other models do not**

- ANN이 DT보다 결과가 안 좋게 나온 이유를 고민해야 함

- numerical이 적을 시 ANN의 효율이 떨어짐

- 이 numerical의 경우도 값이 다양하지 않음

- categorical만 이용해서 한 것도 비교(프로젝트에 새로운 dmm 자료 추가하기)

**• Compare Business Intelligence models with lift charts over a specific class value only on test set**

- 기본 Default 설정으로 진행하면 됨

- score 값을 비교하는 것도 좋은 방법

- 이 모델을 기반으로 Population 20%을 target 했을 때, 84% correctly predicted 가능함

**• Compare Business Intelligence models against a random model and the ideal model on lift charts**

- Random과 ideal에 대한 기본적인 설명으로 발표 진행(수업 때 배운 내용 그대로)

- 그 후 NB, DT, ANN의 결과를 살펴보자. 특정 population의 lift를 비교

**• Present Profit chart based on the best Business Intelligence model**

- Marketing 비용, 이윤 등등을 추산해야 할 듯

- Revenue의 경우는 Marketing의 성공 시 예금 받을 수 있는 금액(평균값을 구할 필요가 있을 듯)

- VC의 경우는 전화 비용이므로 사실상 거의 들지 않는다고 보면 된다. 혹은 인건비로 가정하면 된다. 이 경우에는 비정규직을 사용해도 무방하기에 인건비를 이에 맞춰서 적용해도 될 것 같다. 또한 통화비도 VC에 포함해야 한다.

- FC를 Outsourcing

**• List three most interesting patterns or rules from the best Business Intelligence model for your classification**

- 참고로 DT의 결과값을 가지고만 작업하면 됨

**3) Other tasks**

**• Present the goal of your clustering, association rule or sequence clustering task**

**• Interpret and present any interesting outputs that you believe decision makers should know**

- Macro factor가 많이 사라짐 -> Personal한 게 더 많이 반영

- Class variable에 yes를 표명한 cluster의 특징 중에 눈에 띄는 것을 파악하는 게 좋을 듯

- Clustering 결과를 보면 Age가 높은 cluster가 Campaign 지표가 낮음. 근데 특이하게도 Y 값은 제일 높게 나옴(나이가 많을수록 저축의 여유가 생긴다?)

-> classification에서는 Age에 대한 변수가 덜 강조가 되었는데 이에 대해서도 고려해보는 것도 좋은 방법

-> 물론 굉장히 위험한 Assumption일 수 있으니 조원들과의 상의가 반드시 필요할 듯

**4) Lessons learned**

**• Present any things that you like to share**

- Dependency network에서 가장 높은 값이 나왔던 결과물을 기반으로

- Business, Finance적인 reason을 붙이도록 하기

- October의 경우는 3분기의 가장 마지막 경우, 3월의 경우도 1분기의 마지막 날이라는 것을 명심해야 한다. 또한 ANN에서 보면 각 분기의 마지막 달에 효과적임을 알 수 있다

-> 분기가 끝나면 돈이 많이 생겨서?

-> 이 때에 집중적으로 Telemarketing을 진행하는 것이 좋은 방법일 수 있음

- 저축은 경기침체로 연결됨(인과의 선후를 따질 필요가 있을 듯)

-> 경기침체를 증명하는 지표로 #employed가 맞지 않을까 싶다

- Client attribute의 경우 다른 두 종류의 attribute보다 연관성이 떨어지는 것을 발견할 수 있다.

-> Client 관련 attribute 중에서는 default만 연관성을 보였음(반대로 보면 나머지 client attribute들은 연관성이 많이 떨어진다는 것을 암시)

-> 이 지표는 앞으로 마케팅을 할 때 덜 고려해도 된다는 것을 암시(이러한 조사도 search cost에 포함되기에 그것을 줄일 수 있는 하나의 방법이라고 볼 수 있음)

- Euribor rate는 사실상 interest rate와 같다고 봐도 되는데, 이 값이 낮아지는데도 저축이 늘어나는 이유를 분석할 필요가 있음(이 부분의 경우는 논문을 참조)

-> 경기가 침체되면 이자율이 떨어지는가?

-> 경기가 침체되면 실제로 저축이 늘어나는가?

- duration이 가장 중요한 변수

-> Previous call과 관련된 부분이 중요. 고객 관리적인 측면(?)

- Attribute들을 보면 marketing 담당자가 control할 수 없는 사회적인 변수가 있지만, control할 수 있는 변수들도 분명히 존재한다