**Instructions**

This is a group project. A maximum of 4 students can form a group.

Pick a data mining problem of your choice. You can find so many datasets on Kaggle or if you have contacts (acquaintances and friends) in the local industry then you may collect data from a company as well. In the case of Kaggle, the selected dataset must be of significant size (at least in tens of thousands). However, in the case of local industry data, please get approval from me if the size is small.

You need to submit a report in which you need to follow the complete CRISP-DM methodology by‚ identifying the key (data mining related)‚ question(s)/challenge(s), describing the data set and after application of relevant analytics technique, describing your findings in detail.

Don't focus too much attention on optimizing the evaluation metric rather pay attention to the utility/value of your findings and how it would benefit the concerned organization. Also, identify the limitations of your study and what future data collection suggestions you would have for the organization.

To prepare the report, please revisit the CRISP-DM material covered in the course (or online) and build your report around those steps. As a top-level guideline, your report must contain:

1) Problem Description: This includes the background of the problem and the key data mining technique that you aim to apply.   
2) Data Description: Size of the dataset and description of the columns, Quality of the data (missing values proportion, constant columns, etc.)  
3) Data Pre-Processing: What steps did you take to prepare the data.  
4) Model Building + Evaluation: Which models did you try, which turned out to be the best (based on which metric), etc. How long did it take for a single evaluation? The description of the machine(s) on which you ran your experiments. Which software/libraries did you use, etc? Note: Don't submit/apply anything which was not studied in the course. So for instance, if we haven't studied deep learning in this course then don't submit a deep learning based analysis. The purpose of the project is to judge how you have grasped the concepts introduced in this course.  
5) Findings: This must include the insight you got from the data set, the limitations of your study, what advice would you like to give to the organization in terms of future data collection and processes, etc. If the organization plans to implement your solution, what are the key points related to deployment? When would the model expire, etc?