**Data Mining Group Project  
Page will be updated frequently. Please check for latest version.**

The project is an opportunity for you to explore a machine learning related problem of your choice. The idea is to perform an end-to-end machine-learning project on a real-world data set. You will go through the entire machine-learning life cycle – data collection, cleaning, exploratory data analysis, applying machine-learning models, and evaluation of the models.

You will work in teams of **three to four** people. Please follow the key dates and deliverable described below. Submit your interim reports to d2l, and on a team web page.

**Project Proposal (Due: March 21)**

Your project proposal (1-2 pages) should include:

* Project title, team name, team information, and a public web page for your project.
* Project description: description of the problem, why it is interesting, and who benefits.
* Project goals: what you plan to achieve, and how you would evaluate your model.
* Data Set – description of data set, how are you collecting the data? You may benefit by looking at the Data Set repositories posted under “Useful Links” section of the course web page.
* Tools – specific tools/package that you plan to use for this.
* Literature Review – Collect 2-5 papers or resources related to your problem/area and summarize them. Type in your keywords and search for related papers/ industry white papers using your favorite search engine. Go to ACM (<https://dl.acm.org/>) and IEEE Digital libraries (<https://www.computer.org/csdl>) .

**Midway Report and Short Presentation (Due: April 10)**

This should be a 3-4 pages short report and it serves as a checkpoint. This will help you make sure you are on right track. Your report should highlight what you have accomplished so far, and plans for the rest of the semester. Make sure to update your team web page that reflects your progress such as your mid-term report, link to your GitHub page, etc.

Your midway report should be a template of your final report. You may build on your project proposal. The report should have paper title, team information, abstract, introduction (which typically includes problem statement, motivation), related work, data set and features, methods (models), results and discussion, and then conclusions. Some of these sections will be incomplete, and that is expected.

Be sure to include each team member’s contribution in the appendix of the report.

Upload 3-5 slides describing your progress to the shareable google slide deck given here.

<https://docs.google.com/presentation/d/1wlUzfdUcQ6iYfhmdfIDGV-5hZ72U7w3E3unnAMIm6Ng/edit?usp=sharing>

You will present a quick summary (3-5 minutes per team) on **April 12**. The schedule of presentations will follow the order given below.

|  |  |  |
| --- | --- | --- |
| 1 | German Credit Worthiness <https://github.com/ViditKalani/MSCS-5610-Data-Mining-Final-Project/upload/master> | Shoun Abraham, Vidit Kalani |
| 2 | Iowa Liquor Sales <https://mail929.github.io/MU-COSC4610-Project> | Liam Fruzyna, Katy Weathington, and Sarah Graupman |
| 3 | Steamed Data | Jake Kearns |
| 4 | SanTran  <https://bkupz.github.io/SanTran-WebPage/> | Ben Walczak, Brandon Kupczyk, Charlie Morley, Shivam Thakrar, and Shivani Kohli |
| 5 | An Analysis of Home Field Advantage  <http://www.mscs.mu.edu/~akosla/> | Alex Kosla, Nina Lasswell, Bria Powell, and Franco Reda |
| 6 | Predicting Rising NBA Stars  <https://github.com/kevinetta/nba_data_science> | Kevin Etta, Joe Marotta, and Chandan Matta |
| 7 | A Data Mining Classification Approach to Predict Marquette’s Computing Program Candidate Matriculation  <http://www.cloud-ag.net/ananta/> | Imran Reza Ananta, and Badrun Nahar Rumpa |
| 8 | Student Loan Prediction and Repayment Risk  <https://teamlosdatos.wixsite.com/studentloan> | Alfredo Antolinez, Misti Stevens, and Omar Waller |
| 9 | Forecasting a Data Center’s Environment with Weather Data  <https://sites.google.com/view/marq-dm-s18-data-centers> | Kenyon Mitchell, and Brett Bodenburg |
| 10 | Machine Learning in Bank Marketing | Shikan Zheng, Zhi Du, Yihan Cao, and Hui Jiao |
| 11 | Predicting Seasonal Player Offensive Metrics in Baseball  <https://agattone2.wixsite.com/thebaseballboys> | Alex Gattone, Scott Coyne, Rene Mercado, and Patrick J. McGee |
| 12 | Bank Deposit | Wenliang Hu, Zixun Zhang, and Miaomiao Zhao |
| 13 | One Step Towards Movie Review Analysis for Movie Rating Prediction  <https://pnitu.github.io/textmining/> | Matthew Shafis, Nihel Charfi, Paromita Nitu, and Zachary Boyd |
| 14 | Auto Selector  <https://github.com/rani009/Team-Torque> | Hsiaoan Wang, Priyanka Annapureddy, and Rani Sebastian |

**Final Project Presentations (May 1, May 3rd)**

Upload 6-8 slides describing your final presentation to the shareable google slide deck given here. Each team will have 8 -10 minutes to present their work.

<https://docs.google.com/presentation/d/1fblj0pj9l5tmAvynHXaYBOzeNBrcuqWe_bsEY2MX4v4/edit?usp=sharing>

**Final Report (Due: May 8)**

The final report is a complete version of your midway report. Submit your final report and associated code.

Some sample reports from Stanford’s offering of Data Mining course are given here for your reference. [[sample report 1](SampleProjectReports-1.pdf)] [[sample report 2](SampleProjectReports-2.pdf)] [[sample report 3](SampleProjectReports-3.pdf)] [[sample report 4](SampleProjectReports-4.pdf)].

Highlight where relevant any technical designs and challenging code accomplishments. Include your project legacy discussing what you have accomplished, tying the concepts learned in the class to your project and the skills you have acquired independently from the instructor and classroom discussion. Also, include what you would like to accomplish in future, given more time for this project. Be sure to include each team member’s contribution in the appendix of the report.