PEKING UNIVERSITY, GUANGHUA SCHOOL OF MANAGEMENT
Machine Learning and Artificial Intelligence

Group Project

Prof. Yingjie Zhang

Fall 2021

1 COURSE PROJECT GUIDELINES

This course project offers an opportunity for you to explore an interesting and practical machine learning problem in the context of a real-world business-related dataset. A typical project consists of defining an interesting business-related question and formulating it as a machine learning task, picking a real-world dataset, applying one or more appropriate and well-known machine learning techniques as baselines, and extending these baselines in creative and interesting ways.

Projects can be done in teams of **at most 5** students. Team members are responsible for dividing up the work equally and making sure that each member contributes. The course project will be worth 20% of your final grade. In this document, we describe the detailed requirements in completing this course project, as well as some suggestions of choosing ML-related datasets and questions.

2 DELIVERABLES

This course project requires 3 delivarables:

- **Proposal** (Due date: October 10, 2021, 1 page)
- **Class Presentation** (November 25, 2021)
- **Final Report** (Due date: December 9, 2021)

2.1 PROPOSAL

A **one-page** proposal is due on **October 10, 2021**. You have to submit your proposal via Blackboard. Only one copy per group is required. You are encouraged to discuss your ideas with the instructor or TA before submitting the proposal.

A complete proposal should include the following information:

- Project title
- Project descriptions (e.g., why it is an interesting question in the real world)
- Dataset (including both sources and brief description, such as data fields and summary statistics)
- Teammates and work division
- Potential methodologies (i.e., discuss the ML algorithms you plan to use, including **at least one supervised and at least one unsupervised learning techniques**)

2.2 Class Presentation

All students should present their project in class. Detailed instructions will be posted later.

- *Slides* – The quality of the slides. Do they get the points across? Are they appealing and attention getting? Do they facilitate an effective presentation?
- *Presentation* – The quality of the delivered presentation. How were the points delivered? Was the communication persuasive and informative? Did the tone grab and keep the audience's attention? Did the content flow well? Were the speakers effective in delivering the content?
- *Content* – Was the question they aimed to solve interesting? Were the methodologies they chose proper? Were the analyses process accurate and correct? Did the content reflect an understanding of the topic and a high degree of analysis?
- *Questions* – How did the group handle questions? Were they prepared with good answers? Did they handle the questioners with respect and with an eye to convincing the audience of their views?
- *Extra* – Did the presenters do something "extra" that caught the attention of the audience or kept the audience entertained and interested?

2.3 Final Report

The final report is due on **December 9, 2021** without page limits. A complete final report should include:

- A detailed description of the entire project (e.g., project description, dataset description, methodologies, results, and implications)
- Your python codes (with comments)

3 Grading

Some general tips:

- higher grades will be given to projects that analyze bigger datasets compared to small ones
- higher grades will be given to projects that explore multiple up-to-date machine learning models
- higher grades will be given to individuals who provide critical insights to other teams (members) during the Q&A process

Academic Integrity: All work (including results, discussions, and codes) you submit should be created by you and should be an original representation of your work.

4 Suggestions

Note: This is a challenging and also open-ended project, with no pre-defined “correct” answer. It is up to you to locate a dataset (or any new trend) that is interesting and possible to analyze in a meaningful manner. Projects that combine multiple datasets and multiple ML-related methodologies will a receive higher grade.

Here are links to several interesting datasets:

- Urban Computing
- NYC Open Data
- Yahoo! Research Data
- Movie Data
- World Bank Open Data
- Basketball Data
- NFL Data
- Google Trends
- Dataset repository organized by UC Irvine
- Tianchi
- Kaggle