**MIE1413 Project Proposal**

Yunjie Xu 1008601951

**Project Target:**

Based on the probability to predict the future each policyholder from the MEB insurance company is "High value" or "Low value."

**Background Information:**

I am an actuary at MEB Insurance and have been asked to assist the marketing department. They have been collecting data on past policyholders. For each policyholder, they have determined if they are “high value” or “low value” based on profitability. They would like to be able to predict for each prospective policyholder if they will be high or low value1.

**Dataset Introduction:**

1. Number of Observations: 48842
2. Number of Variables: 8
3. Dependent Variable: value\_flag (category variable)
4. Independent Variable: score, hours\_per\_week, cap\_gain, occupation, marital\_status, education\_num, age.

**Model Planning:**

1. I am considering doing the GLM logistics regression to fit the data to form the best model. I may apply model checking, feature selections, ridge or lasso, and other techniques during these processes.
2. The classification tree may also be an excellent way to approach the best prediction model.

**Data Source:**

This dataset from SOA Exam PA December 13, Test Project. Original Source: <https://www.soa.org/education/exam-req/syllabus-study-materials/edu-multiple-choice-exam/>

**Reference:**

[1] <https://www.soa.org/globalassets/assets/files/edu/2019/spring/solutions/2019-06-14-exam-pa-project-statement.pdf>