**Capstone Project Proposal**

**Mentor**: Eric Rynerson

**Mentee**: Lou Zhang

**Date**: 04/27/2016

**Predicting Survey Submissions**

The goal of this project is to create a predictive model of whether or not companies will submit data to a proprietary monthly survey conducted by a manufacturing association. This survey has been conducted this survey since 1997 and has about 700k rows of data, detailing each participant’s monthly submission and various attributes of that submission. This will be coupled with other data sources both public and private, and the merged dataset will ultimately list detail such as (on a month by month basis):

1. Whether or not the company submitted data for that month
2. What month they submitted for
3. When they submitted the data in reference to when the data period opened
4. How long they have been in the program
5. What values the company submitted
6. Which products those values correspond to
   1. Share of Census total for the NAICS category the survey captures
7. Geography of company
8. Size of company
9. Engagement with Trade Shows

This will be coupled with macroeconomic data such as

1. Industrial production
2. Manufacturing technology production

The deliverable is a predictive model of whether or not each participant will submit/not submit in the upcoming month. A “prediction score” from 0 to 1 will be attributed to each participant. A cutoff of 0.5 will serve as the minimum benchmark for probable submission. This prediction score will be used to determine if a pre-emptive participation reminder will need to be sent to those companies, which will help increase participation and provide better customer service.

**Client**

A manufacturing association is the client for this project. This association represents over 75% (by dollar volume) of manufacturing technology industry in the United States.. This association administers the survey being analyzed, which is the largest survey of manufacturing technology production in the United States.

**Data**

Survey Proprietary Data, Federal Reserve Economic Database, American Survey of Manufacturers. All proprietary data sets readily available in company databases.

**Approach**

·        Collect and merge all the data - this will require significant wrangling.

·        Scope the problem down to make it both more tractable, as well as relevant.

·        Model it as a classification problem to predict the probability of submission for each company. Train the model on an earlier data set and test it on later subsets.

·        Evaluate effectiveness of the model by tracking it against historical data.

·        Tune model based on accuracy to historical data

·        Revise model and test on current data

**Deliverables**

Code, Report, Slide Deck