**Final Capstone – Predicting Payments on Medical Procedures**

1. What is the problem you are attempting to solve?

Using 5 years of payment data for health care procedures, can we predict payments better than the current "Group Average" methodology being used?

1. How is your solution valuable?

How much insurance companies and other government entities pay for various procedures is somewhat of a black box. However, to forecast future cash flows, an estimate of the total charges for each claim. Thus, a more precise estimate can improve future operational and financial planning, and valuation exercises. The following questions may be addressed with a better prediction:

1. How many new employees for next year?
2. How much can we invest back into the business?
3. How much should we set aside for future tax payments?
4. What is the current valuation of our company?
5. How is your solution valuable?
6. What is your data source and how will you access it?

A national healthcare provider with hundreds of sites, performing thousands of procedures from 2014 – 2018. The data does not include any personal identifying information, and I will not reveal any specific names around insurance companies, payers, or individual claims.

1. What techniques from the course do you anticipate using?
2. SQL to organize and group the data
3. Visualization in Python
4. Supervised Learning Regressors to predict payment amounts
5. What do you anticipate to be the biggest challenge you’ll face?

The challenges I anticipate most include:

1. Setting up the supervised learning exercise since the data is not grouped at the claim/procedure level, and spans 4 tables
2. Creating new columns to summarize and group data to aid in visualization
3. The dataset is 22 million records with over 40 columns, and some of the categorical variables have thousands of responses. Thus, it could cause memory errors in Python, and require a lot of resources to train various models. I will look to do feature selection, and perhaps consolidate categories where it makes sense to mitigate this.
4. Setting up the exercise will require some thought since time is a consideration. Since I’ll compare the current method used to a newer model, I must also simulate how a current prediction can only be made using data from the past.