Capstone Project Proposal

Medication Recommender

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
  The problem I am trying to solve is the amount of time that doctors take in giving certain patients a drug that can potentially save them.
- What industry/realm/domain does this apply to?
  This would apply to the health industry. It would also apply to the pharmaceutical industry.
- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)
  I wanted to do this project for my undergrad but did not have the opportunity to do it. Now that I have the chance to dig into it, I want to complete it. Its also very aligned with my current project at the NIAID. I see how researchers and doctors constantly have problems with the data they work with.

Data Understanding

- What data will you collect?
  The data that I will be collecting is bioactivity data.
- Is there a plan for how to get the data (API request, direct download, etc.)?
  The plan for how to get the data is by using the ChEMBL Database.
- Are the features that will be used described clearly?
  Yes, the features that I will be using are clear and well documented.

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?
  The preprocessing steps will be mostly in the realm of data cleaning. I will also use the Lipinski Descriptor calculation. This calculation will tell me drug likeness of compounds
- What are some of the cleaning/pre-processing challenges for this data?
  Some challenges may arise in having uniformly distributed data. Uneven distribution of data points may be an issue as well.

Modeling

- What modeling techniques are most appropriate for your problem?
  Regression and Random Forest seems to be the most adequate.
- What is your target variable? (remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)
  My target variable will be the protein thrombopoietin receptor.

- Is this a regression or classification problem?
  This is a classification problem but will use some regression as well.

Evaluation

- What metrics will you use to determine success (MAE, RMSE, etc.)
  One of the important metrics is the Mann-Whitney U metrics test. This will help me look at the difference between the two bioactivity classes. It tests for statistical significance in the classes. I also used the MSE score.

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?
  I plan on using random forest