The project is about a new cycle for the model where:

1. The requirement analysis will be revisited and validated with stockholders (you can use your imagination to capture the stakeholders’ requirements)
   1. We reviewed the current plan and requirements and determined the following
      1. The original pre-processing did not clean out as much bad data as it should
      2. The original data was too granular in being able to predict accurate outcomes
      3. The model did not seem appropriate based on our research
   2. We proposed that the intent of the project is determine if a review was positive or negative, and suggested that a much higher accuracy could be obtained if the granularity of prediction was decreased.
2. Formulate the specification from the requirements
   1. The new requirements will involve predicting positive/negative reviews based on a text string
      * 1. Signoff with client
   2. The initial input for training the model can still use the scaled review (0-1 on .1 increments)
      1. The training will map 0-.4 to negative and .6-1 as positive
         1. Signoff with client
   3. Accuracy must be above 80%
      * 1. Signoff with client
3. Put in an architecture for the system
   1. Azure ML will be used to train and predict
   2. The Azure ML web service integration will be used to handle the API for submitting review test and return a positive or negative result
      1. i.e. a twitter feed could be used to get real time analysis feedback on sentiment of the movie
4. Plan a remodeling for the system
   1. Map 11 categories (0,.1,.2,…,.9,1) to negative/positive
   2. Use custom python code version controlled in Git
   3. Explore pre-processing text to clean up input
   4. Explore different feature extraction from text
      1. LDA, Word2Vec, Tf-Idf, Bag-Of-words
         1. Research indicated - Tf-Idf was best
   5. Explore different modelling related to prediction
      1. Random Forest, Naïve Bayes and Logistic regression were common in the literature
5. Implement the new model
   1. Developed pre-processing based on literature and reviewing the data
      1. Python code
   2. Used Tf-ldf in python code for feature extraction
      1. Test with at 500 and 1000
   3. Tested the following machine learning models
      1. Logistic Regression
      2. Random Forest
6. Verify the model’s outputs against the requirements
   1. Outcomes ???
7. Document the process step by step
   1. Initial work tried to evaluate existing model and get it to work with existing requirements
   2. We could not come to good predictions with such granular predictions and took a new approach to discuss with client an alternative approach
   3. Worked on various text pre-processing to clean out questionable data
   4. Implemented the Tf-Idf vectorization to get output to use in Logistic Regression and Random Forest
8. Manage the process step by step
   1. Trello
   2. Git
   3. Twice weekly meetings
   4. Initial research on Azure ML, text processing, text vectorization and modelling in Azure space
9. Communicate the results step by step
   1. Git and Trello
      1. Shared some Trello stories with client for review
         1. Web Service API
            1. Text input
            2. Positive/Negative response
         2. Time frame for reviewing results of work
10. Deliver the product in a presentation
    1. Have slides of model running (input and output API)
       1. “this movie bad” – negative
       2. “great move” – positive
    2. We’ll figure out the slides for presentation and discuss who presents what