Project Proposal

Yuzhong Huang, Wilson Tang, Jay Woo

* In a couple of paragraphs, describe the key ideas of your proposed project? What is your MVP? What are your stretch goals?

The health IT space has exploded in recent years with multiple innovative technical solutions aiming to improve everything from reducing readmission rates to optimizing cancer treatment. Many of these solutions take advantage of the vast amounts of healthcare data available to help create helpful tools for chealthcare professionals and patients. Our team is aiming to use electronic health record data to first create a predictive model using neural networks and EHR data, then use it to create a diagnostic tool. Our project is intended to give us more experience/learn about the application of neural networks to healthcare data, working on reorganizing data to, working on creating well presented data visualizations and tools, and work on an real-world, impactful data project.

Our MVP is to create a project website that describe our project, the limitations of our model with visualizations of the relationships identified, and a link to our diagnostic tool. Our stretch goals would be to integrate our diagnostic tool to utilize HL7/FHIR standards which are used in the healthcare space and make the tool potentially integratable with actual EHR systems.

* To the best of your current knowledge, what datasets will you use for your project? Are there any obstacles you foresee in terms of getting access to the data?

We will be using electronic health record data, found from a Kaggle competition held in 2012. The data is technically HIPAA compliant, so we won’t have to deal with any legal issues. If we find that this dataset is not what we want, we could potentially get data from some external collaborators who have expressed interest in our project.

* What are the most important new skills / techniques you will have to learn to be successful in this project? If you think some of these skills would be useful for us to cover in class, please indicate which ones.

We would have to learn new machine learning algorithms to be able to make predictions using our data. More specifically, we are interested in applying a neural network with feature extraction. We may also have to look into the naming protocols for the dataset, which could taking a lot of digging into external resources. Other than that, we might also need to learn how to build a website that can host our diagnostic tool, if we decide to make our project available on the Internet.

\*I think that neural networks both basic and advanced can be a good topic to cover

* Outline a rough timeline for the major milestones of your project. This will mainly be useful to refer back to as we move through the project.

Organizing and Exploring Dataset: 1.5 weeks

Neural Network: 1.5 weeks

Feature learning: 2 weeks

Diagnosis Tool and website building: 1 weeks

* What do you view as the biggest risks to you being successful on this project?

Getting overwhelmed with the amount of potential information given in the dataset and understanding all the re-codings the dataset offers.

Scoping the project appropriately because each major milestone is intensive and can possibly end up taking up a lot more time than initially allocated.

* Given each of your YOGAs (see [here](https://sites.google.com/site/datascience16/home/assignments/project-3#TOC-YOGA)), in what ways is this project well-aligned with these goals, and in what ways is it misaligned? If there are ways in which it is not well-aligned, please provide a potential strategy for bringing the project and your learning goals into better alignment.

Our YOGAs are pretty well aligned, in that we are all interested in learning some new algorithm and presenting our results in a compelling manner. Some of our goals need to be modified (ex. applying knowledge of SigSys to the project), but we are all willing to change at least one of our goals to ensure that we all get the most out of our work.

* Jay
  + Learn a new machine learning algorithm
    - This goal fits pretty well and probably won’t need any changing
  + Tell a compelling story through data visualizations
    - We adjusted our project slightly to include a data presentation portion (what trends did we notice while developing our neural network tool, for instance)
  + Incorporate concepts learned from other classes
    - I will probably need to change this goal, mainly because I don’t really see any opportunities to accomplish this with the given dataset. This goal will probably change to something about data exploration, which I struggled with a lot for the previous project.
* Wilson
  + Apply Neural Networks to Healthcare Data
    - Fits well, we plan on using neural networks as our algorithm
  + Work on presenting work and creating helpful data visualizations
    - We adjusted and plan on making sure to well-document the project and present trends our model identifies within the website
  + Work on practical and impactful project
    - Our project is aiming to solve a real need and is based on actual data + health standards
* Yuzhong
  + Feature learning
    - This goal fits well with this project since we will build an autoencoder to represent the data using sparse matrix
  + Presentation
    - This goal also fits well with the project. In the end of our project, we need to present our research result. So I could get chance to practice my presentation skills
  + Deep learning
    - Our project relies on the deep learning model we built, so this goal fit well too.