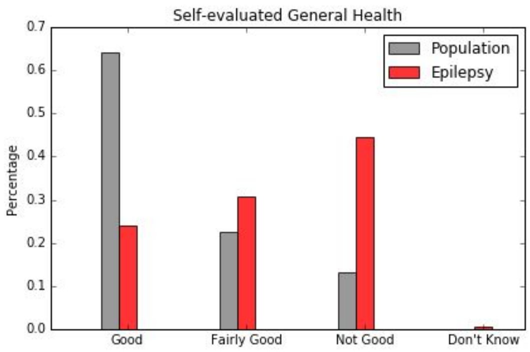
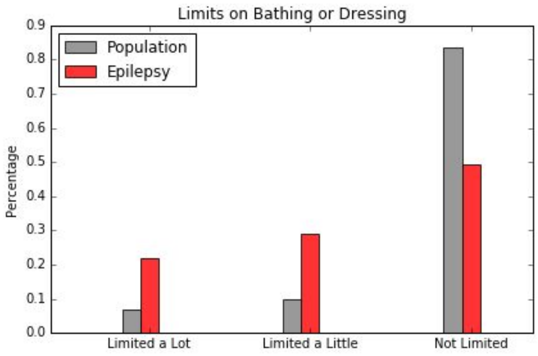
Seizure Prediction Milestones 4 and 5
Xu Si, Hung-Yi Wu, and James Zatsiorsky

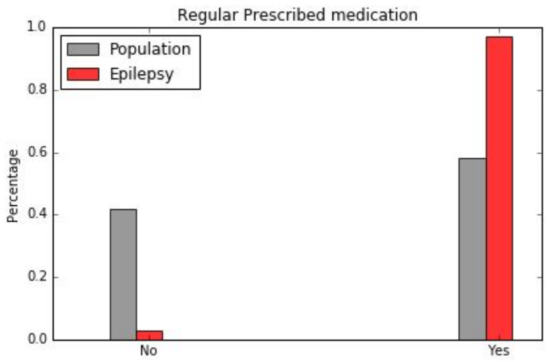
Milestone 4

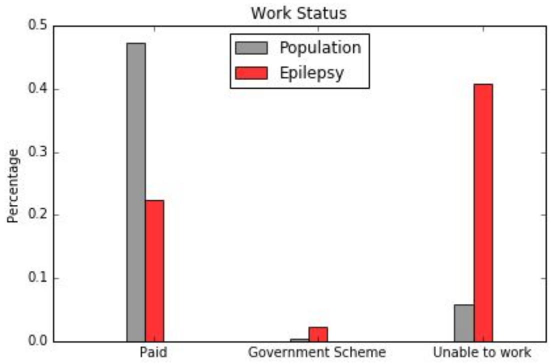
Since the last milestone, we focused on three main tasks. We present them below, with brief descriptions of what we noticed. Python code is provided with our submission.

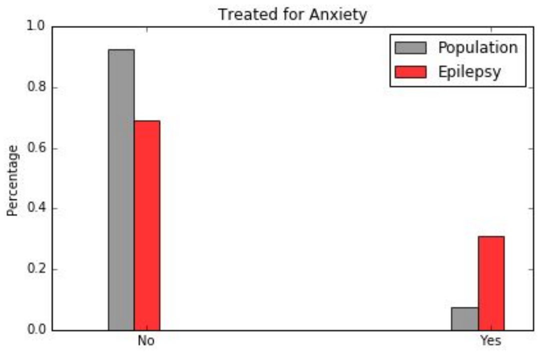
1. **Life of epilepsy patients**
   To better our storytelling, we want to start from the life of epilepsy patients. Are they suffering? What do they think about life? Are they living a normal life? For this part, we used data from Welsh Health Status (detailed analysis is in the ipython notebook). And find surprising results. They have less positive attitudes toward life and may suffer from anxiety and depression. Some are even permanently unable to work. Several of the findings are displayed below:

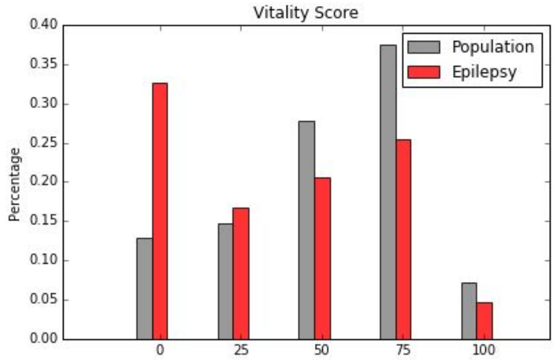












Are they beaten by epilepsy? What are people living with epilepsy saying? We want to take a look at their life. Then we go to the blogs of epilepsy patients and find their courage and positive attitude!

2. **Additional data wrangling and cleanup for epilepsy prediction**

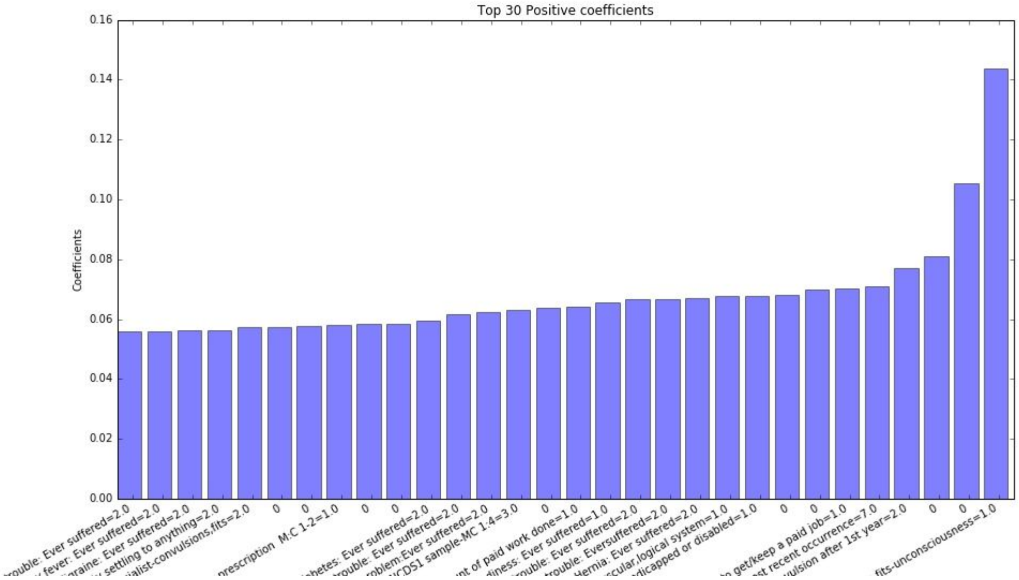
    - **Choose predictors:** The dataset we have (NCDS) is really messy. There are a huge amount of predictors: some are potential predictors that may cause epilepsy/seizures and some are the results of having epilepsy (e.g. reason for prescription and occurrence of latest convolution). Then I decided to roughly filter out predictors from e.g. medical questionnaire and parental questionnaire (detailed info is in the ipython notebooks) that could be cause for having epilepsy.

    - **Choose datasets and response:** Another decision we make is to merge two of the longitude NCSD survey. We combines sweep 0-3 (at age 7,11,16) and sweep 5 (at age 33) and use the epilepsy indicator from sweep 5 as our response. In total we have 141 epilepsy patients and 11,163 “healthy” patients. The reason we merge the two datasets is that we want to include more predictors and consider that some patients do not show epileptic syndrome until grow up into adults.

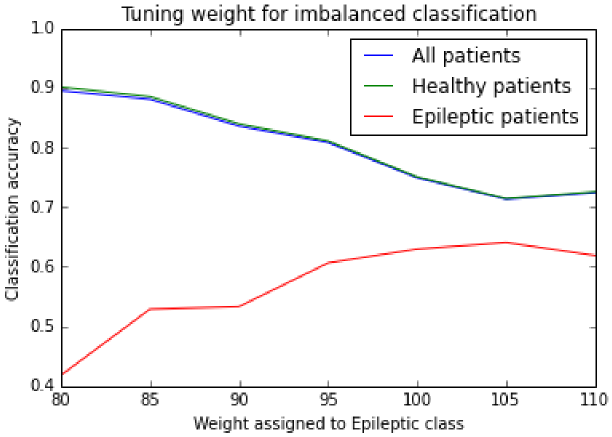
    - **Missing values:** After choosing our predictors and response, we end up having 1449 predictors with one response. Nearly all predictors contain missing values, and some predictors have more than half blanks. For numeric predictors, we fill with the mean and for categorical predictors we try to either treat NaN as a new category or fill with the mode.

3. **Logistic regression**
    We first try to fit a logistic regression and get accuracy of around 73% on both classes (with epilepsy or not). Then we sort the predictors by coefficients and got the top 20

predictors.



4. **Random Forest**
   We fit a random forest classification model to see how accurate of a classifier we could build for patients with and without epilepsy, and to examine the features of importance in the dataset. Due to the imbalance of the two classes - there are 11,163 patients without epilepsy and 141 patients with epilepsy in our dataset - we tuned the class-weight parameter of the sklearn RandomForestClassifier constructor to find an optimal tradeoff between accuracies on each class. Below and on the left we show a chart demonstrating the effect of tuning this weight on the cross-validated classification accuracy. Increasing the weight assigned to the epileptic class tends to increase the accuracy on epileptic patients and decrease the accuracy on healthy patients. We found that a weight of ~105 provided relatively high accuracies for each class.



   After tuning some of the other parameters of the random forest model, we built a classifier that classified epileptic patients with 73% accuracy and healthy patients with 73% accuracy. Some of the top predictors of interest for this model are "Psychiatric,psychological treatment", "Fit or convulsion after 1st year", "General health: _4.0", "No. times driven/ridden over limit in last 12 months: _4.0." We will dive deeper into the top predictors as part of our future work.

Milestone 5

We have several key areas of interest that we plan to focus on in the days leading up to the project deadline. We list them below with descriptions on how we plan to tackle them.

- **Website design**: Our website is mainly divided in three parts.
    1. First, introduction to seizures and epilepsy. Visualize the life of epilepsy patients.
    2. Epilepsy prediction. We will do both prediction for normal people and people who has first seizures. Build models and analysis accuracy.
    3. Evaluation. Test the probability/whether a person will have epilepsy by filling a form.
- **Model selection and tuning**

    1. **Decide on Performance Metric**
       As exemplified by the weight-tuning in the random forest classification plot in Milestone 4, changing the class-weight used in our models can have a substantial effect on the classification rates for epileptic and healthy patients. For example, in the case of our random forest classification, setting the epileptic class weight close to 1 assigns a ~100% accuracy rate to classifying healthy patients but a very low accuracy rate to classifying epileptic patients. With the final tuned random forest model we developed, we had a 73% accuracy rate on each class. We should formally develop what we mean by a "good" classifier, and then tune the parameters to our model to optimize that performance metric. We will further analysis **confusion matrix** and **ROC curves** and also try to do more accurate feature selection/
    2. **Remove Relevant Predictors**
       Although we worked to remove any predictors that directly related to our response variable, i.e. some predictor which directly states whether a patient has epilepsy, we should comb through the important predictors found by each of our models and make sure that we remove any that mention epilepsy. We don't want these predictors in our model because new patients won't have this information - they're wondering whether they have epilepsy!
    3. **Look at Effect of Having a First Seizure**
       We've built classifiers for whether a patient has epilepsy. We would also like to build a classifier for the presence of epilepsy *given that the patient has had a first seizure*.
    4. **Build a Classifier for New Patient Data**
       As a part of our website, we are interested in exploring the possibility of adding an epilepsy classifier with new patient data. A user can input some data (to be determined) and a "diagnosis" will be returned. This might be best for users that

have already experienced a first seizure, since the number of predictors required to predict epilepsy might be fewer conditional on having a first seizure. We will use our results from bullet point 3 above to determine the best way to go about doing this.