Checkpoint 2

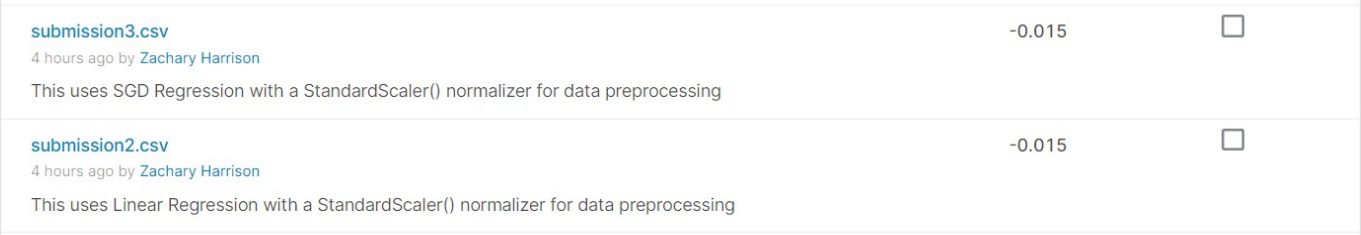
This will be a one-page report that describes what we've done so far, such as descriptive statistics about the dataset, any data preprocessing or feature engineering conducted, the comparison results of different approaches we've run, and our plan for the rest of the semester.

Truth

- Team members: Zachary Harrison and Keldon Boehmer

Submissions

We have made two submissions using Linear Regression and Logistic Regression. They went rather poorly as they performed worse than our dummy submission, but that is largely due to our poor data preprocessing and the competition's data set being rancid. Here are the submissions:



submission3.csv
4 hours ago by Zachary Harrison
This uses SGD Regression with a StandardScaler() normalizer for data preprocessing
-0.015 ☐

submission2.csv
4 hours ago by Zachary Harrison
This uses Linear Regression with a StandardScaler() normalizer for data preprocessing
-0.015 ☐

Now, even though these submissions somehow achieved a negative score (I know we're amazing, but please save your shocked applause for later), we believe there to be some promise in what we have achieved thus far. I encourage you to take a further look at our work on GitHub.

Data Preprocessing:

- The vast majority of this task is going to be data preprocessing. You see, there are only 2 features that are intuitively understood by any Regression of Classification model, and both of these values are constant in all testing samples. The only variable of concern is the protein_sequence variable, which is a sequence of letters.

Feature Selection:

- I'm not sure if it counts as feature selection of data preprocessing, but method by which we modify the protein_sequence variable is extremely important. It is discussed further in the hyperlinks above.

Plan

1. We need to think of a good way to split up the protein_sequence variable so that a model can actually be trained on it. Using its length wasn't a great idea because it was also near-constant for all the test sets.
2. Consider data normalization or even data standardization
3. Research and implement superior regression models
4. Look at submissions/code from other submissions and incorporate their *good* ideas into our submissions