General Assembly DC - DAT5
Predicting HIV Progression in Patients
By: Nana Adu-Krow

1.0 Problem Statement and Hypothesis

By applying our knowledge of viral load and cd4+ counts found in patients with HIV can we actively predict whether that patient would respond well to a 16 week therapy program?

Another possibility I'd like to explore is whether we can effectively use the amino acid sequence in our predictions to see whether patients would respond positively or negatively.

I believe through the ensembling method and combining different models emphasizing various features I should be able to come up with an effective algorithm. With the conceptual approach of combining logistic regression, single tree decision analysis, as well as naive bayes I should be able to accurately predict whether a patient will respond well to treatment or not.

2.0 Description of the Dataset

The data set was provided through one of the older Kaggle Competitions. It highlights the different features given on both the train and test data set. The Dataset is organized as follows:
Col-1: Patient ID
Col-2: Responder status ("1" for patients who improved and "0" otherwise)
Col-3: Protease nucleotide sequence (if available)
Col-4: Reverse Transcriptase nucleotide sequence (if available)
Col-5: Viral load at the beginning of therapy (log-10 units)
Col-6: CD4 count at the beginning of therapy

3.0 Preprocessing

The data was collected from the Kaggle Competition website and appropriately gave two csv files. The first was the training dataset and the second smaller data was the testing dataset.

4.0 Exploring the Data

The first thing that I was interested in doing was looking at the training data set and seeing the visual relationship between the viral load and CD4+ count of the patients who responded well and the patients who didn't respond well. There are definitely more patients who didn't respond well to patients who did, and you can see the negative relationship between viral load and cd4 count in patients with responder status 0. Patients with responder status 1 had a bit more of positive slope in comparison.

However with what I want to do with prediction, looking at linear regression relationships wouldn't help. Classification methods would be the method I use moving forward.

5.0 Choosing the Features
Trying to find patterns in the amino acid sequence was actually much harder than I thought. With that in mind I moved forward with using the two remaining features in the dataset. Based on reading light research material on HIV and how it affects the CD4+ count and Viral Load of patients I felt I would be able to find a relevant pattern for machine learning purposes.

6.0 Modeling Process
I decided to run logistic regressions on the two features in three different ways. The first feature I wanted to look at was the Viral Load. After running the model on the training set I was able to receive a predictive score of 79.6% and the Area under the Curve was 74.7%. What was interesting was that the CD4+ count did slightly better with an 80% predictive score but significantly worse with the Area under the Curve score of 64%. Combining the two features did worse than focusing on the Viral Load mainly because it was dragged down by the CD4+ count predictor.

When I applied the model with the Viral Load towards the test data set I got an 80% prediction score and 74.4% Area under the Curve.

7.0 Challenges and Success
What I came across that was most challenging was the Protease Sequence and the Reverse Transcriptase Sequence. Trying to find patterns within those columns proved to be very complicated. I tried to select the most common 4 amino acid combinations using CountVectorizer but the results weren't really useable. I could get some patterns but it didn't translate well to the test data.

I was successfully able to use logistic regression on the two features I decided to focus on. It was also good to look at the Area Under the Curve to highlight which features worked better than others. Looking at the AUC allowed me to see that using the feature of Viral Load was a much better predictor of seeing how someone would respond to treatment.

8.0 Possible Extensions and/or Applications
This area of data science is the future of healthcare and treatment in this country. By understanding a patients DNA sequence or other varying feature sets (i.e., a person's Viral Load and CD4+ count), treatment can be customized towards an individual. This approach could ultimately save funds for hospitals that give expensive catch all treatments as well as empower patients to selecting the best possible options for them.

9.0
Some of the key takeaways from this project is the importance of learning different evaluation metrics. The competition asked for misclassification error and I couldn't properly apply that code to my work. This shows me that there's more practice and work I need to undertake for my own personal journey into python and statistics.

Understanding Area under the Curve was wholly beneficial in picking the correct feature to run the Logistic Regression on. Had I used the feature that had the best accuracy, my score would have been significantly worse especially if I used it on the test data.