Predicting HIV Progression

Using Data to Predict HIV Prognosis

Nana Adu-Krow

Agenda

- Hypothesis
- Background
- Data Analysis
- Visualizations
- Moving Forward
- Questions?

Hypothesis

- Is it possible to predict how severe HIV progression will continue based on past data?

Background

When monitoring the progress of a HIV infection, its important to use tests that give quantitative results.

Two main indicators used to measure HIV progression

1. HIV Viral Load - Number of viral particles in 1 mL of blood. The higher the VL count is the more active the immune system is.
2. CD4+ Cell Count - Approximation of white blood cells in 1 mL of blood. The higher the CD4 count is the seemingly healthier the subject.

Data Analysis

In our training data we have 1000 patients.

6 columns on each patient.

- Patient ID
- Responder Status
- Protease Nucleotide Sequence
- Reverse Transcriptase Nucleotide Sequence
- Viral Load
- CD4+ Count

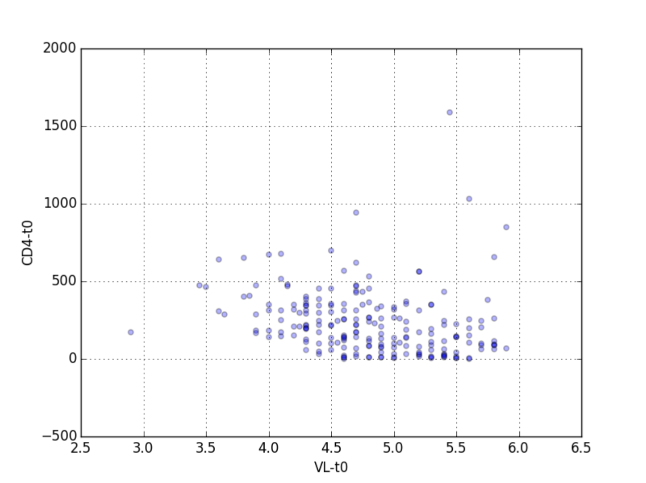
We want to focus on the Response Status and VL and CD4+ Count.

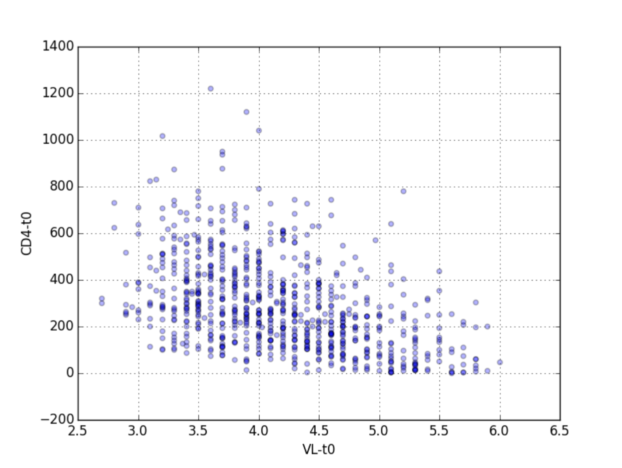
Pre-Processing Steps

This information and dataset was found from the Kaggle website.

It has since been closed but users are able to freely utilize the data.

Data Visualization





Feature Selection

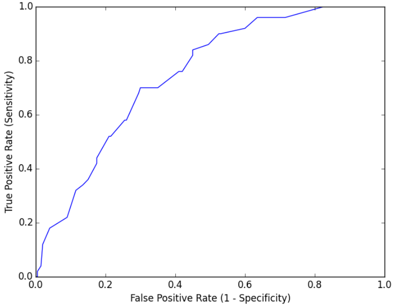
CD4+ Count and VL were the only features I was able to find a pattern.

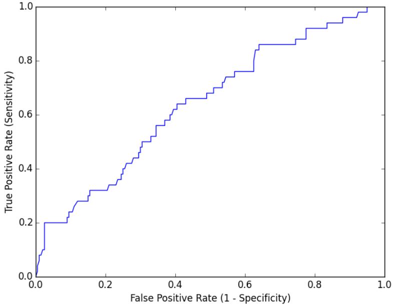
Modeling Process

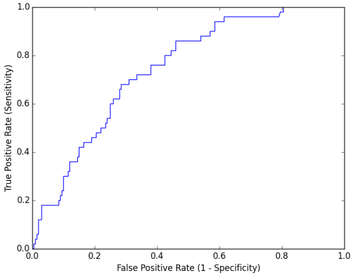
Logistic Regression

- Viral Load
  - Pred: 79.6%
  - Auc: 74.7%
  - 10 fold validation (Auc): 76.2%
- CD4+ Count
  - Predi: 80%
  - Auc: 64%
  - 10 fold validation (Auc): 60.2%
- VL & CD4+ Count
  - Pred: 79.6%
  - Auc: 74.4%
  - 10 fold validation (Auc): 75.8%

Area Under the Curve







Testing on VL feature

When testing on the test data set I received an accuracy of 79.2% with AUC of 74.4%!

Challenges

Amino acid sequences are complicated...

Key Learnings

- I have a lot to learn about when it comes to Data Science…
  - The competition submissions asked for misclassification error method which I improperly tried to use Stack Overflow for!
- Accuracy isn’t always the best metric for how a model might do outside the sample.

Potential Applications

Questions???