**FINAL PROJECT REPORT– HANH NGUYEN**

­­BIOS 648 HIGH DIMENSION DATA ANALYSIS

**1. Background**

In this project, regression problem will be solved by using appropriate statistical models on HIV data. In this project, the outcomes, which are the log viral load and CD4 count, would be separately predicted based on the baseline data by the end of year 1 and by the end of year 2. In this project, I will use linear regression with subset selection method, random forest and support vector machine to predict the outcomes of interest. The prediction power of these methods will be compared based on their regression standard error in the training period and testing period. For random forest and linear regression model, important features used for prediction will be extracted for qualitatively assessing the intuitive reasonableness of the model.

**2. Description of Data**

The predictors used for prediction includes: log of viral load at time 0 (lrna0), CD4 count at time 0 (cd40), the drug used for the patient (called “arm” – which is a factor of 6 levels presenting 6 therapeutic drug combination), and the data for mutation status which is presented under a binary type of data (1 = mutated and 0 = wild-type) at the first 240 codons of the reverse transcriptase (RT) region and the first 99 codons of the protease (PR) region.

Some important findings have been found after exploring the data:

* The number of patients under each “arm” are quite uniform among 6 groups

A picture containing computer

Description automatically generated

Figure Arm distribution

* The mean value of lrna0, lrna1, lrna2 (log viral load at time 0, year 1, year 2) and cd40, cd41, cd42 (CD4 count at time 0, year 1 and year 2) are also quite similar among 6 arms
* The distribution of lrna1, lrna2, cd40, cd41 and cd42 are skewed to the right. This should be considered seriously when using linear model.

A close up of a map

Description automatically generated

Figure Distributions of lrna and cd4 at time 0, year 1 and year 2

* There are 47 NA found in the comprehensive data (including dat384, rtmut0 and prmut0)
* There are 71 mutation positions of both the reverse transcriptase and the protease found to be constant in all patients (71 predictors have zero variance)

**3. Statistical Methods**

There are three models to be implemented: random forest, linear regression with subset selection of variables, support vector machine (SVM) and generalized linear model with elastic net regularization. Each method requires different data pre-processing methods.

**3.1 Exploring data**

**3.2 Dealing with missing value**

**A screenshot of a cell phone

Description automatically generated**

Figure Missing value distribution

**3.3 Pre-processing data**

**3.4 Fitting models**

**3.4.1 Random forest**

**3.4.2 Linear model with backward subset selection**

**3.4.3 Support vector machine**

**3.4.4 Generalized linear model with elastic net regularization**

**4. Results**

**4.1 Prediction of lrna1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Random forest | Linear model with backward subset selection | Support vector machine | GLM with elastic net regularization |
| Train error |  |  |  |  |
| Test error |  |  |  |  |
| 95% CI |  |  |  |  |

**4.2 Prediction of lrna2**

**4.3 Prediction of cd41**

**4.4 Prediction of cd42**

**5. Conclusions**