How to classify Alzheimer’s disease and identify those presymptomatic individuals with mild cognitive impairment (MCI) who will eventually convert to Alzheimer’s is the issue addressed in Sandip Ray’s manuscript. Statistical methods including significance analysis of microarrays (SAM), unsupervised clustering algorithm, predictive analysis of microarrays (PAM, 10-fold cross-validation regularization and logistic regression) were used in Ray’s work. His conclusion is that 18 signaling proteins in blood plasma are found that can be used to classify blinded samples from Alzheimer’s and control subjects with close to 90% accuracy. Also, these signaling proteins can help identify patients who had MCI that progressed to Alzheimer’s disease 2-6 years later. Biological analysis of the 18 proteins points to systemic dysregulation of hematopoiesis, immune responses, apoptosis and neuronal support in presymptomatic Alzheimer’s disease, which makes the conclusion more convincing.

The first and most interesting part of Ray’s work for me is that whenever he draws a conclusion, he wants to verify it from other perspectives. For instance, after SAM identified 19 proteins with highly significant differences in expression between Alzheimer’s and NDC samples, he applied an unsupervised clustering algorithm based on the similarity in abundance of these 19 markers which produced two main clusters that contained mostly Alzheimer’s or NDC samples respectively. Similarly, after PAM identified 18 predictors and classified Alzheimer’s and NDC samples with very high accuracy, the unsupervised clustering based on these 18 markers was able to separate Alzheimer’s and NDC samples at the same time. From high level, this manuscript is not merely about statistical analysis but combined with the biological one. Moreover, they are consistent with each other.

Another impressive part is that these 18 predictors also perform well in distinguishing Alzheimer’s samples from other neurological diseases and rheumatoid arthritis. Though this may not be Ray’s initial intention, the result does support his statement that these 18 biomarkers certainly form an Alzheimer’s-specific signature.

To go further, we may want to find the reason why there exists a discrepancy between SAM and PAM —— the missing biomarker CCL22. Why it shows significant difference but is discarded by PAM? What will the regression model and prediction error be with 19 predictors? What’s the correlation between CCL22 and M-CSF? CCL22 and M-CSF are the nearest in the phylogenetic tree, maybe CCL22 is somewhat collinear with M-CSF. By solving these questions, we may be able to explain the missing CCL22 thus makes the work more persuasive.

To make predictions regarding post-college earnings and debt of alumni by machine learning models is the issue discussed in Monica’s project. Techniques like single regression imputation, sequential forward-based feature selection, linear/locally weighted linear/K-nearest neighbors/support vector regression and neural networks were applied during the process. The incremental model selection process indicated that regression imputation of privacy-suppressed values improved performance of models. Besides, weighted linear regression outperformed other models with below 10% and 18% average error for earning and debt data respectively. Moreover, this model performed well towards law schools though law schools made up only little of the data.

One of the most attractive highlights in Monica’s project is privacy-suppressed values handling. Instead of simply removing all features with any privacy-suppressed entries or merely setting missing values to the mean of observed values, single regression imputation was applied. To make the regression model statistically meaningful, they imposed a requirement that imputed features must have missing data for less than 30% of schools. Since privacy-suppressed values occurred in potentially useful metrics, the aforementioned implementation does increase the credibility of the model.

Model competing is a fascinating part in this project as well. There are totally 5 classes of optimized models among which best weighted linear regression model outperformed others. The authors discussed the superiority of the best weighted linear regression by exploring whether the model generalized well outside their dataset and got a positive result. Besides, they also analyzed the reason why support vector regression did much worse than all other models. Due to the reasonable steps that they gone through model refining, one could safely draw the conclusion that the best model among the 5 was indeed the best weighted linear regression.

Although there is no doubt that this project is amazing with huge workload, there are still some deficiencies as far as I am concerned. The very first one is the representativeness of the dataset the authors selected. Their goal was to make prediction of post-collegiate earnings and debt while the data they collected was from post-college earnings and debt of alumni who were on federal financial aid. Students who received federal financial aid may share some unique features that potentially influenced their earnings and debt. Without eliminating this possibility, the authors couldn’t simply assume that their data is representative. I do know the fact that there are some cases when it is extreme hard to retrieve data, but I strongly recommend obtaining data from all kinds of alumni.

The second one could be the poor performance of their best model given data of trade schools. As the author stated, the current algorithm has trouble extending to such schools, which indicates that future earnings and debt might be best characterized by a different set of features.

For future work, partnering with the U.S. Department of Education to gain access to all kinds of data (from all kinds of alumni and maybe some privacy-suppressed data) would be a big plus. More generalized models for imputation can be involved to handle the remaining privacy-suppressed data better if given enough computational resource. Finally, improving the performance towards trade schools by possibly selecting new features could form the project an organic whole.