**Introduction**

Regularized regression techniques have been shown to decrease variance, albeit for small tradeoff in the increase of bias. This is accomplished by including penalization terms on the parameter estimates of the model. In linear regression, estimates can become inflated due to issues with collinearity, and as such, can increase the SSE of the model. In an attempt to reduce the SSE, penalization terms are introduced which control the magnitude of the estimates. Two such penalization terms are the L1 and L2 penalties. The Elastic Net model incorporates both terms and has the advantage in that it enables effective regularization while also allowing for a feature selection quality.

In penalized regression it is often necessary to center and scale the data. Centering is necessary in order to avoid the use of an intercept term. Likewise, given that Elastic Net puts constraints on the size of the coefficients associated to each variable, this value will depend on the variable itself. Therefore, scaling is also necessary.

In genomic data, continuous variables such as distances are often heavily skewed. Therefore, log2 transformations are typically used in order to better normalize the data. It is unclear if, or how much more, beneficial it is to log2 transform the data before centering and scaling when performing a regularized regression model. Here we aim to test the predictive ability of an elastic net model with and without a log2 tranformation vs with and without standardization of continuous predictors.

**Methods**

Publicly available topologically associating domain data was obtained from GEO with accession GSE53525. The domain data was in the form of two-column genomic coordinates from the GM12878 cell line, with each coordinate representing a TAD boundary. The coordinates were concatenated into a long vector and were sorted. The genome was then binned into a series of 1kb intervals (500 bases on either side of the boundary point) and an indicator vector Y was created based on whether there was a TAD boundary in the 1kb interval (Y=1) or not (Y=0). There were 16,700 TAD boundaries out of 2766314 possible 1kb intervals. Therefore, the class probabilities were given as: 0.994 for Y=0 and 0.006 for Y=1.

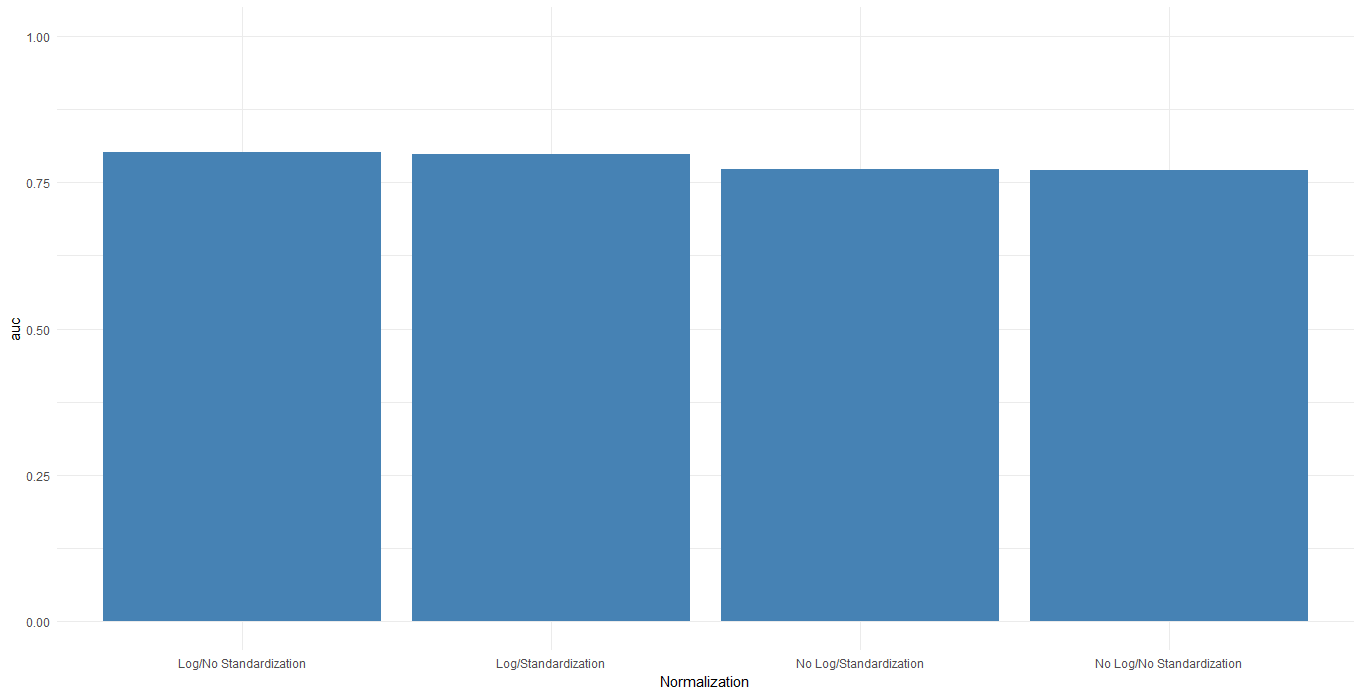
Various genomic annotations were obtained from the GM12878 cell line of the ENCODE Project. These annotations made up the feature space **X**, including p predictor variables ({X₁,...,X­p}). Two types of predictors were created for each annotation. First, a binary variable was created denoting whether or not the 1kb bin overlapped with the annotation or not. Second, a continuous variable was calculated for each annotation as the distance to the nearest coordinate from the 1kb bin. The total feature space consisted of 103 predictors.

A total of four elastic net models were performed including: with log transformation and standardization, without log transformation and with standardization, with log transformation and without standardization, and without log transformation and without standardization. Ten bootstrap samples were taken from the majority class to create a balanced classification data, while five fold cross validation, repeated 3 times was used in order to tune for the L1 and L2 penalties. Model performance was evaluated using ROC curves and corresponding AUCs. Likewise we looked at the feature selection quality of the models. All analysis was perfomed in R version 3.4.2.

**Results**

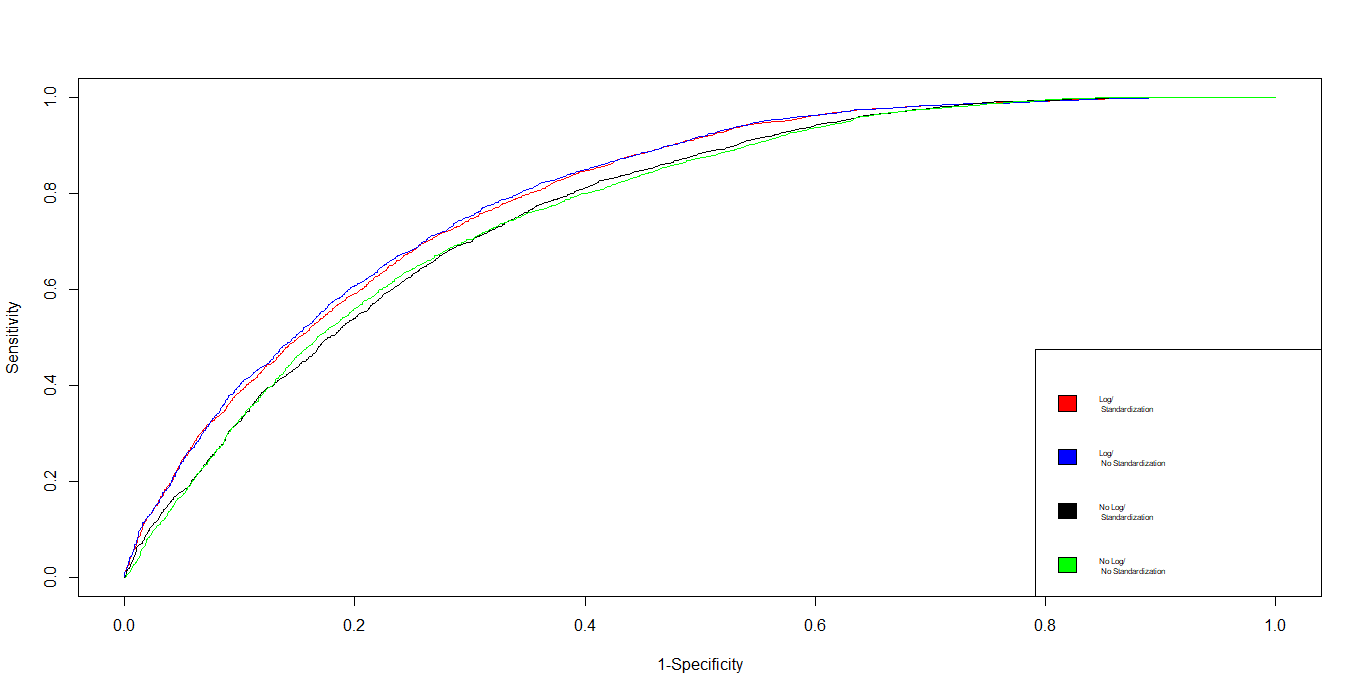
From Figure 1 we can see that the model with the largest AUC was the elastic net that implemented a log2 transformation, with no standardization. The model with a log2 transformation as well as standardization was not far behind (AUCs of 0.802 and 0.799 respectively). The corresponding ROC curves are provided for each combination in Figure 2. This reaffirmed the similar performance between the standardized and nonstandardized models including log2 transformations.

|  |  |
| --- | --- |
| **Elastic Net Model Performance** | |
| **SMOTE Combination** | **AUC** |
| Log and No Standardization | 0.802 |
| Log and Standardization | 0.799 |
| No Log and Standardization | 0.772 |
| No Log and No Standardization | 0.771 |

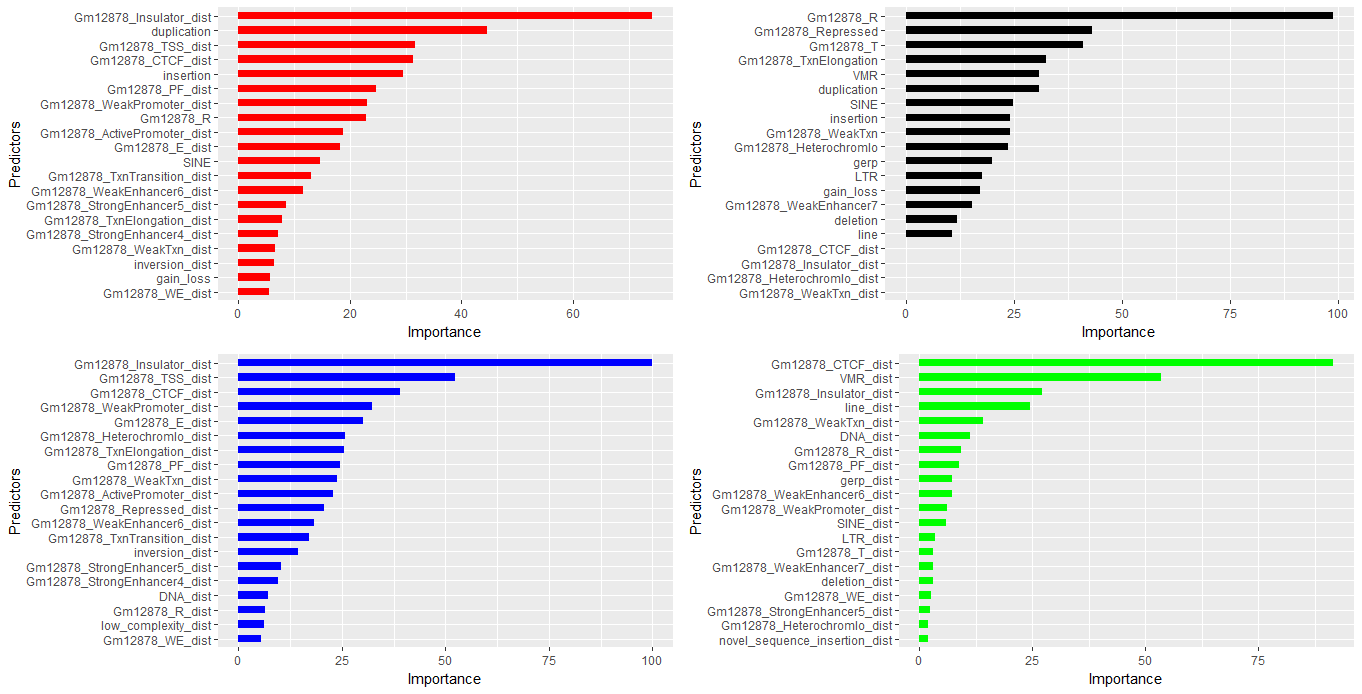


**Figure 1.** Model performance metrics in the form of AUCs for each of the 4 combinations of models involving either a log2 transform or standardized covariates.

Figure 3 provides the feature importance of the four models. The first column are the two models with a log transform, with (red) and without (blue) standardized covariates. Each model includes the variable insulator and CTCF distances in the top 4 features which is to be expected. The models in the second column, without log transform show conflicting results. The model without a log transformation but with standardization has the insulator or CTCF variables near the bottom of the top 20 important features.



**Figure 2.** ROC curves for each of the 4 combinations of models involving either a log2 transform or standardized covariates. Log transform with standardization is in red, log transform



**Figure 3.** Variable importance plots for each of the 4 models. Top left: with log transformation and standardization; top right: without log transformation and with standardization; bottom left: with log transformation and no standardization; bottom right: without log transformation and without standardization.

**Discussion**

When working with penalized regression models it is good practice to make sure continuous variable are normalized as well as on the same scale. This has significant effects in both performance and interpretability. In our study, we saw that the models with log transformation out performed the two models without it. Within the two log transformed models, the ROC curves and corresponding AUCs were very similar. Likewise, the feature importance plots presented the insulator and CTCF variables as top predictive features which were to be expected. This confirms the need for at least implementing a log2 transformation. Even though, the model without the standardization performed slightly better, it should be noted that only 10 bootstrap samples were performed when creating balanced data. Therefore, it is likely that when implementing more bootstraps (closer to 1000) the results using standardized covariates will be asymptotically similar. Thus, we conclude that using a log2 transformation as well as standardized covariates is the best approach here.