**Introduction**

High-throughput chromatin conformation capture (Hi-C) has allowed scientists to gain a better understanding of the spatial organization of the human genome. One of the most important elements of this organization is known as topologically associating domains (TADs). These TADs are formed when DNA loops around and comes into contact with itself. It is through this manner that the genome is able to fit compactly within the nucleous of a cell.

It has been determined that this looping structure plays a key role in gene regulation by drawing transcription factor binding sites near genes. It has been reported that the DNA binding protein CTCF, as well as other insulator proteins, are strongly associated with the formation of DNA loops. Important questions remain, however, such as determining what other genomic factors are associated with TAD formation as well as devising an efficient model for determining said factors.

Here, we aim to identify the molecular drivers that are most predictive of the topological arrangements of chromatin. Using TAD boundary locations, machine learning algorithms were tuned in order to accurately classify whether or not a TAD boundary was located in a genomic bin. A classic and often used model for this scenario is the logistic regression model. While this method has its merits, it lacks an ability to both handle possibly correlated predictors and identify which predictors are, indeed, most predictive. Because of this we have chosen to employ a variety of machine learning techniques including the elastic net, random forest, and gradient boosting machine.

Likewise, given the sparcity of topological domains throughout the genome, we tend to have heavily imbalanced classifications. It has been shown that a classification model with imbalanced classes will perform poorly. Therefore, we employ two methods to remedy this. The first method consists of taking multiple bootstrap samples from the majority class to create a balanced sample. The second method utilizes the SMOTE command from the DMwR package in R. Each method is then evaluated and compared to the baseline logistic regression model.

**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, 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.

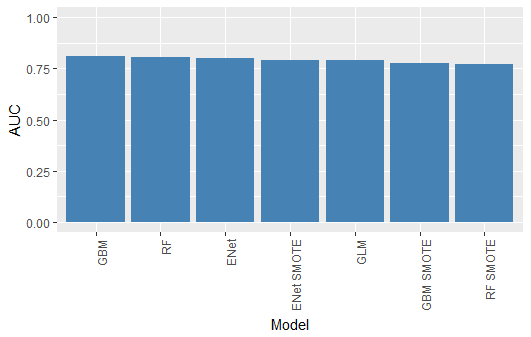
Various genomic annotations were obtained from the GM12878 cell line of the ENCODE Project. These annotations made up the feature space **X**, including p predictor variables ({X₁,...,X­p}). Two types of predictors were created for each annotation. First, a binary variable was created denoting whether 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 preliminary logistic regression model was performed and used as a baseline. Prior to model implementation, all predictors with near zero variance were omitted. Also, any variable that was a linear combination of another was removed. The final data set consisted of 61 predictors.

In order to create balanced classes, 5 bootstrap samples were taken from the majority class and combined with the minority class. The model performance metrics were then averaged across each sample. Seperately, the SMOTE command from the DMwR package was implemented. SMOTE stands for Synthetic Minority Over-sampling TEchnique. Both majority undersampling and minority oversampling is used to create balanced data. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors (here, k=5). Elastic net, random forest, and gbm models were performed for a total of 6 models (3 from concatenated bootstraps and 3 from using SMOTE). For each model, 10-fold cross validation repeated 5 times was considered. Model performance was evaluated using AUCs. Intracomparisons of the three newly implemented models (within each sampling technique) was done by comparing the variable importance rankings. All analysis was perfomed in R version 3.4.2

**Results**

The AUCs for each of the baseline logistic regression, elastic net, random forest, and gbm models (including both sampling techniques) are presented in Figure 1. The baseline logistic regression had an AUC of 0.7882. We see that the baseline logistic regression model under-performs compared to each of the three bootstrapped models. The bootstrapped GBM model performed best with an AUC of 0.8087.



|  |  |  |
| --- | --- | --- |
| **Sampling Technique** | **Model** | **AUC** |
| None | GLM | 0.7882 |
| Bootstrap | Elastic Net | 0.7986 |
|  | Random Foret | 0.8064 |
|  | GBM | 0.8087 |
| SMOTE | Elastic Net | 0.7918 |
|  | Random Foret | 0.7726 |
|  | GBM | 0.7761 |

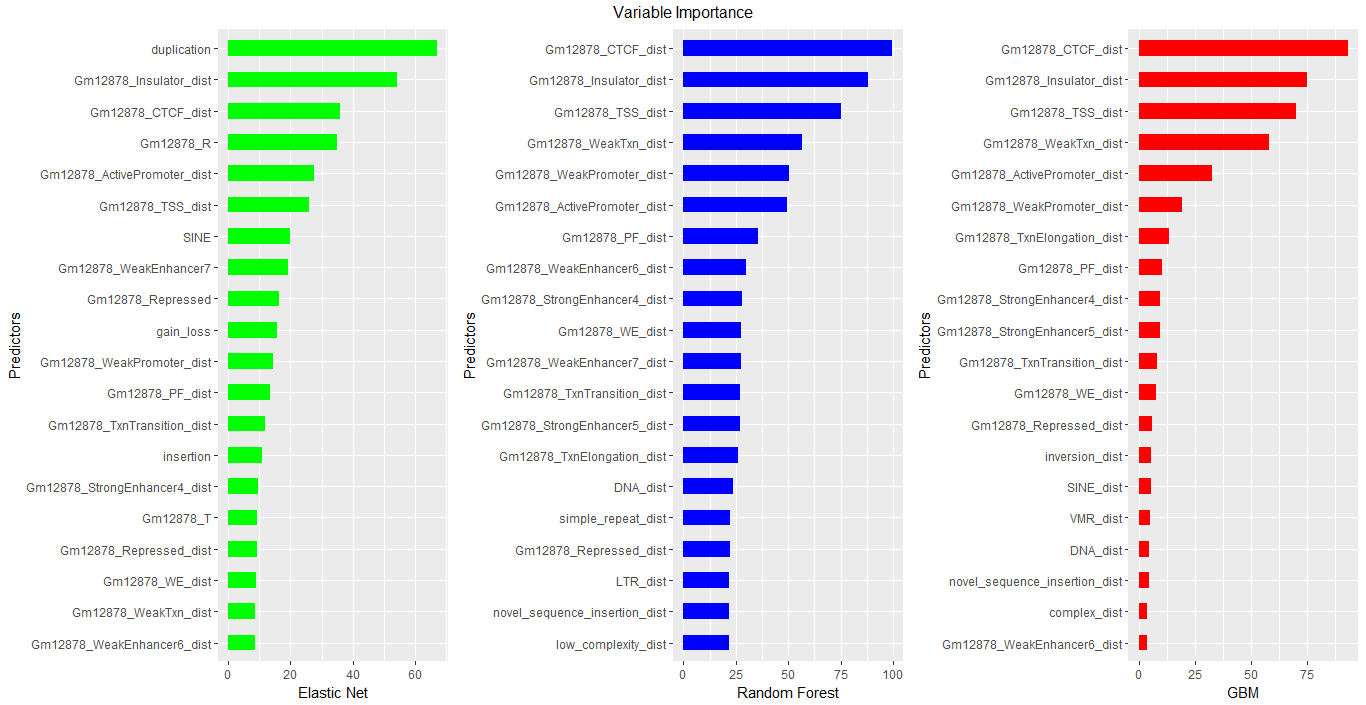
**Figure 1.** Performance metric of each model in the form of AUCs. The bootstrap sampled models performed resoundingly better than both the baseline logistic regression model and the models using SMOTE.

The benefit of using these particular machine learning algorithms is to be able to determine the predictive strength of each feature. Figure 2 & 3 present the variable importance plots of the top 20 variables for each of the two sampling techniques. We see that for the bootstrap sampled date there is heavy agreement between all three models, especially between the random forest and gbm. Both the CTCF and insulator annotation variables have been identified in the top 3 most important variables for classifying TAD boundaries. Tables 1 & 2 present the list of common features between each of the three models as well as their respective rankings. Out of the 20 most important variables, 12 of them were common between all three models for the bootstrap technique. We can see the similarity between the random forest and gbm models by inspecting the rankings. The top 4 most important variables identified were the same in both models. There is some disparity between the elastic net model however. We see there is less of an agreement between each of the three models when using SMOTE. Only 9 out of the 20 top variables were similarly identified in all three models. Likewise, the rankings appear to be less similar among the three.

**Discussion**

Given the sparsity of the data and the multitude of possible predictors that are associated with the location of TAD boundaries, it can be concluded that the logistic regression model is not a viable approach here. Two methods were considered when attempting to remedy the class imbalance problem. We first took multiple bootstrap samples from the majority class and concatenated them with the minority class. Performance results were then aggregated according to the mean. Secondly, we considered creating synthetic observations using the SMOTE command. It was found that all three models when using the bootstrap technique performed better, with the gbm model outperforming all others. Likewise, when using the bootstrap technique there was a greater similarity between the identification of important features between the three models.

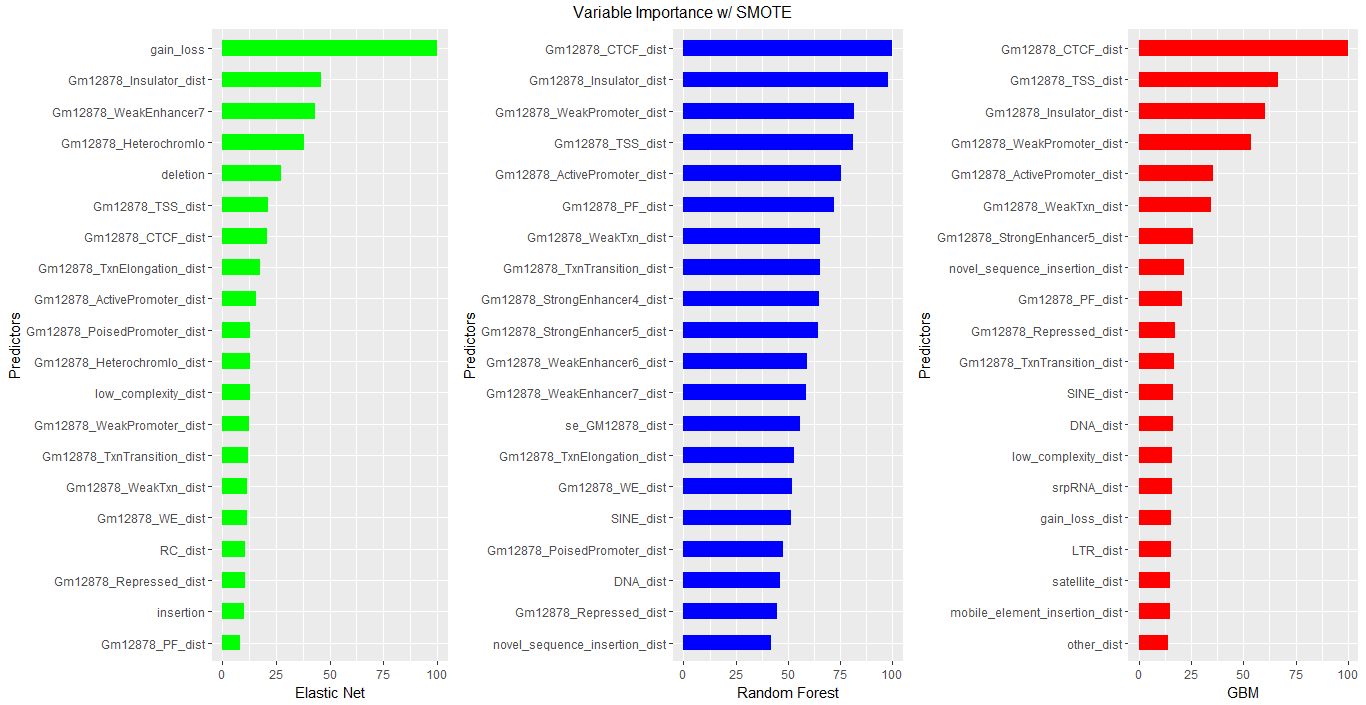
As a limitation, it should be noted that the location of TAD boundaries and genomic annotations was restricted to only chromosome 1. Further analysis is needed to generalize to the full genome. Likewise, only 5 bootstrap samples were considered when creating the balanced classes. The number of bootstrap samples should be increased here to expect for asymptotically stable results.



**Figure 2.** Variable importance plots depicting the top 20 features among the three models elastic net (green), random forest (blue), and gbm (red) when using the bootstrap technique to create balanced classes.

|  |  |  |  |
| --- | --- | --- | --- |
| **Common Features** | | | |
|  | **Model Ranking** | | |
| **Features** | **Elastic-Net** | **Random Forest** | **GBM** |
| Gm12878\_CTCF\_dist | 3 | 1 | 1 |
| Gm12878\_Insulator\_dist | 2 | 2 | 2 |
| Gm12878\_TSS\_dist | 6 | 3 | 3 |
| Gm12878\_WeakTxn\_dist | 19 | 4 | 4 |
| Gm12878\_WeakPromoter\_dist | 11 | 5 | 6 |
| Gm12878\_ActivePromoter\_dist | 5 | 6 | 5 |
| Gm12878\_PF\_dist | 12 | 7 | 8 |
| Gm12878\_WeakEnhancer6\_dist | 20 | 8 | 20 |
| Gm12878\_StrongEnhancer4\_dist | 15 | 9 | 9 |
| Gm12878\_WE\_dist | 18 | 10 | 12 |
| Gm12878\_TxnTransition\_dist | 13 | 12 | 11 |
| Gm12878\_Repressed\_dist | 17 | 17 | 13 |

**Table 1.** Rankings of the commonly identify features among each of the three models when using the boostrap technique.



**Figure 3.** Variable importance plots depicting the top 20 features among the three models elastic net (green), random forest (blue), and gbm (red) when using SMOTE to create balanced classes.

|  |  |  |  |
| --- | --- | --- | --- |
| **Common Features** | | | |
|  | **Model Ranking** | | |
| **Features** | **Elastic-Net** | **Random Forest** | **GBM** |
| Gm12878\_CTCF\_dist | 7 | 1 | 1 |
| Gm12878\_Insulator\_dist | 2 | 2 | 3 |
| Gm12878\_WeakPromoter\_dist | 13 | 3 | 4 |
| Gm12878\_TSS\_dist | 6 | 4 | 2 |
| Gm12878\_ActivePromoter\_dist | 9 | 5 | 5 |
| Gm12878\_PF\_dist | 20 | 6 | 9 |
| Gm12878\_WeakTxn\_dist | 15 | 7 | 6 |
| Gm12878\_TxnTransition\_dist | 14 | 8 | 11 |
| Gm12878\_Repressed\_dist | 18 | 19 | 10 |

**Table 2.** Rankings of the commonly identify features among each of the three models when using SMOTE.