MRR Project 2023
Richard CHEAM & Menghor THUO
10 December 2023

1. Introduction

Recalling the objective of this project is to classify whether a molecule is **toxic** or **non-toxic** to *U2OS* cells. To achieve this, we present two different analytic cases: one involving binary classification using the **initial dataset**, and the other using the **reduced dataset**, with reasoning derived from independent student's t-tests.

2. Collinearity Reduction

Since there exists a strong collinearity between covariates (X), hence, those who are highly correlated will be removed with respect to the result from the first screening of the dataset. To be more concise, only variables that do not have a great statistical impact on classifying the binary target. As a result, we obtained a data frame of **171x391** dimension.

3. Data splitting

In order to prevent the bias from the classification, the data set is randomly split into two different sets: 70% for training (exploration), and the rest is for testing (confirmation).

4. Methodology

For each case of analysis, **5 different methods** are being employed: **full logistic regression**, l_1 and l_2 **penalized logistic regression** (**Lasso** and **Ridge**), **Elastic Net,** and lastly, **K-Nearest Neighbors**. As for the fact that there are a great number of predictors (X) compared to number of observations (n), the base linear regression will not be taken into account. Plus, the penalized techniques will be used in order to shrink the coefficients and make the model much simpler. In addition to that, to make a more reasonable classification without any bias and also to have a cautious decision, for instance, we do not want to have **False Negative Rate**, meaning, we classify a molecule as **non-toxic** when it is actually **toxic** in which it could pose a big issue to *U2OS* cell. By that, Receiver Operating Characteristic (ROC) curve will be taken into account in order to choose a threshold in decision step so that it is balance between **True** and **False Positive Rate**. We then use the chosen **optimal threshold** to make decision.

- For **full logistic model**, due to the fact that this is a binary classification, this method will be the baseline.
- For **Ridge** and **Lasso**, the tuning parameter (λ) is chosen by performing 10 folds cross-validation by selecting the lambda that minimize the error, giving the best accuracy. To this end, a loop of these cross-validations will be run in order to obtain the range of lambda and to see as well how they shrink the coefficients each time.
- For **Elastic Net**, the l_1 and l_2 penalty mixing parameter (α) is chosen from 11 values, 0.1 to 0.9, increment by 0.1 with respect to lambda that minimizes the error rate. We will see as well each time how many explanatory variables remain important.
- Whereas, for **KNN**, to search for optimal k, of Elbow method will be utilized.

Then the metrics utilized in order to evaluate the performance of each model are **misclassification rate**, **accuracy**, **false positive and negative rate,** which they are calculated by using the confusion matrix obtained from the decision step in which the trained models are used to perform binary classification on the test data set.

3.1 Initial data set (collinearity reduced)

For this case, we have a data set of **171x391** dimension: **171** observations, **390** predictors, and **1** binary target, namely, '**Class**'.

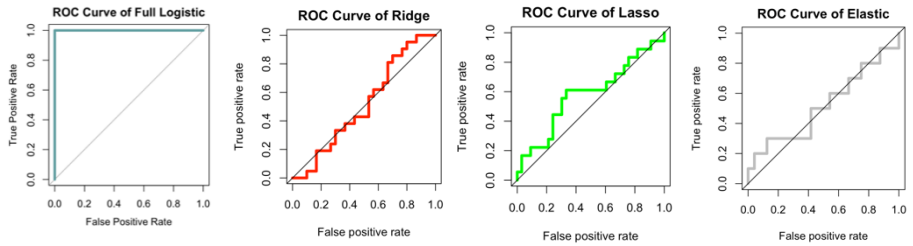

Figure 1: Receiver operating characteristic (ROC) curve for each model (initial dataset)

	Full logistic	Ridge	Lasso	Elastic net
Optimal threshold	0.5	0.1797596	0.04611993	0.3252219
True Positive Rate	100	100	100	66.66667
False Positive Rate	0	83.333333	90	21.739130

Table 1: TPR, FPR of chosen optimal threshold for each model based on the ROC curve (initial dataset)

Visibly, based on the graph above (**Figure 1**), each model has a different shape of curve, however, the penalized ones are moderately identical. We search for the **optimal threshold** in which it gives a reasonable classification. For instance, the best threshold of the ROC curve of **Full Logistic** model would be the point on the top left corner, (0,1), indicating it has 100 **TPR** and 0 **FPR**. Based on the fitting output from R, full logistic model is the result was not reliable as the standard error for each coefficient is too big even after neglecting the high correlated X(s).

	Full logistic	Ridge	Lasso	Elastic net	k-Nearest Neighbors
Lambda		[4.129493, 23.087214] 123.2095 (1se)	[0.05853450, 0.12320952] 0.1232095 (1se) (null)	[0.04567595,0.10072145] 0.1760136	K = 5 (optimal)
Accuracy Rate	53.52941	[49.01961,50.98039]	47.05882	47.05882	62
Error Rate	46.47059	50.98039,49.01961	52.94118	52.94118	38
Predictor	390	390	[7,11]	[8,19]	1203
False Positive Rate	45.86913	83.33333	90	90	29.41176
False Negative Rate	48.37302	0	0	0	56.25
Molecular Descriptor		"SpDiam_Dt", "EE_Dt", "C2SP2", "SpMAD_Dt", "SP.5"	"nHCsats" "EE_Dt" "AATSC7p" "GATS7v" "minssNH" "VR3_Dt" "SHCsats"	"ATSC3e" "nHCsats" "EE_Dt" "SpMin4_Bhi" "maxssNH" "AATSC7p" "MATS7p" "GATS7v"	

Table 2: Performance of each model given the range of lambda by cross-validation and optimal threshold (initial dataset)

By using the optimal threshold above for decision step, we then obtained the information above. **Table 2** shows that the **KNN** has the best accuracy rate, but accuracy is not the only metric to evaluate its performance. Despite having the best classification rate, it has a larger **FNR** compares to the penalized methods. The fact that they have a low accuracy rate is due to the chosen optimal threshold since we would rather make a **FPR** than a **FNR**. Noted as well that each method represents **"EE_Dt"** as the one among the statistical significances.

3.2 Reduced data set

For this case, we have a data set with **171x38** dimension: **171** observations, **37** predictors, and **1** binary target, namely, '**Class**'.

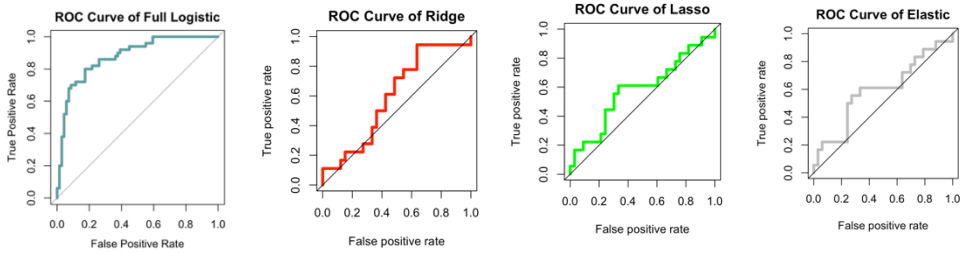

Figure 2: Receiver operating characteristic (ROC) curve for each model (reduced dataset)

	Full logistic	Ridge	Lasso	Elastic net
Optimal threshold	0.2918812	0.28688886	0.334601113	0.33726031
True Positive Rate	92.105263	72.222222	61.111111	61.111111
False Positive Rate	15.853659	48.484848	33.333333	33.333333

Table 3: TPR, FPR of chosen optimal threshold for each model based on the ROC curve (reduced dataset)

According to **Table 3**, in terms of **TPR** and **FPR**, they are all relatively different for the penalized logistic regression models in which the **full logistic** has best metrics, highest in **FPR** and lowest in terms of **FPR**.

	Full logistic	Ridge	Lasso	Elastic net	K-nearest Neighbors
Lambda		[0.75092944,1.19569203] [25.7604,125.2626](1se)	[0.01775565, 0.07866855] 0.1252626 (1se) (null)	[0.02783829,0.05859693] 0.1789465 (1se) (null)	K = 9 (optimal)
Accuracy Rate	72.54902	56.86275	62.7451	64.70588	62
Error Rate	27.45098	43.13725	37.2549	35.29412	38
Predictor	37	37	[2,12]	[9,13]	37
False Positive Rate	18.18182	48.48485	33.33333	33.33333	32.35294
False Negative Rate	44.44444	33.33333	44.44444	38.88889	50
Molecular Descriptor		"C2SP2" "SpMin4_Bhi" "EE_Dt" "SpMin3_Bhe" "ATSC1v" "SpMin3_Bhm"	"minHBint4" "SpMax4_Bhm" "SpMin3_Bhm" "ATSC1v" "nAcid" "EE_Dt" "GATS7v"	"SpMin3_Bhm" "ATSC1v" "nAcid" "EE_Dt" "C2SP2" "GATS7v" "SpMin4_Bhs" "WTPT.1"	

Table 4: Performance of each model given the range of lambda by cross-validation and optimal threshold (reduced dataset)

It highlights **Full Logistic** model as the best with the highest **accuracy rate** and lowest **FPR** but with a strong **FNR**. While **Lasso** and **Elastic Net** beat **Ridge** in terms of **accuracy rate** and **FPR**. Notably, they also shrunk some coefficients to 0 which leads to the further reduction of the predictors to between 10 and 19, except some essential explanatory variables as shown above and once again, **"EE_Dt"** appears an important descriptor.

5. Conclusion

Both two cases of analysis do not really demonstrate a really good classification yet, however, the results are more interpretable when fitting models on the reduced dataset in which the **full logistic** is the best one, but the penalized regression models still at least did comparatively a good job since they produce a simpler model. We would consider trying **Tree-Based** models if we had more time since they handle the large dimensional data well.