IITM-CS5691 : Pattern Recognition and Machine Learning — Release Date: 23 Oct, 2021
New Method Writeup — Due Date: 10 Nov, 2021

Roll No: CS21M026 CS21M009 — Name: Kankan Jana and ASHOK KUMAR THOTA

- Dear Student, You may have tried different methods for predicting each of the clinical descriptors in the data contest. Submit a write-up of the methods chosen for the data contest in the template provided below. **You will have to add the details in your own words and submit it as a team in gradescope**.
- We will run plagiarism checks on codes/write-up, and any detected plagiarism in writing/code will be strictly penalized.

1. (points) [Name of the Method chosen for each descriptor, and its Paradigm: paradigm could be linear/non-linear models, etc.)]:

DESCRIPTER	CLASSIFIER	PARADIGM
C0: 1	Logistic Regression	Linear Model
C0: 2	Logistic Regression	Linear Model
C0: 3	Logistic Regression	Linear Model
C0: 4	Logistic Regression	Linear Model
C0: 5	Logistic Regression	Linear Model
C0: 6	ADABoost	Non-Linear Model

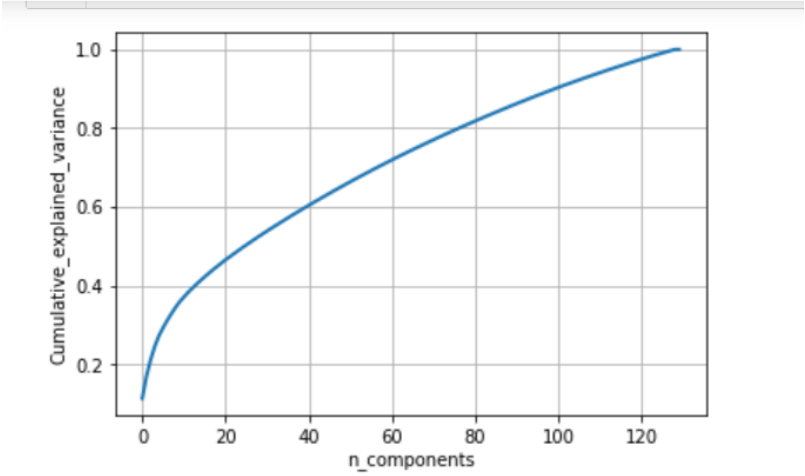
Solution:

2. (points) [Brief description on the dataset: could show graphs illustrating data distribution or brief any additional analysis/data augmentation/data exploration performed]

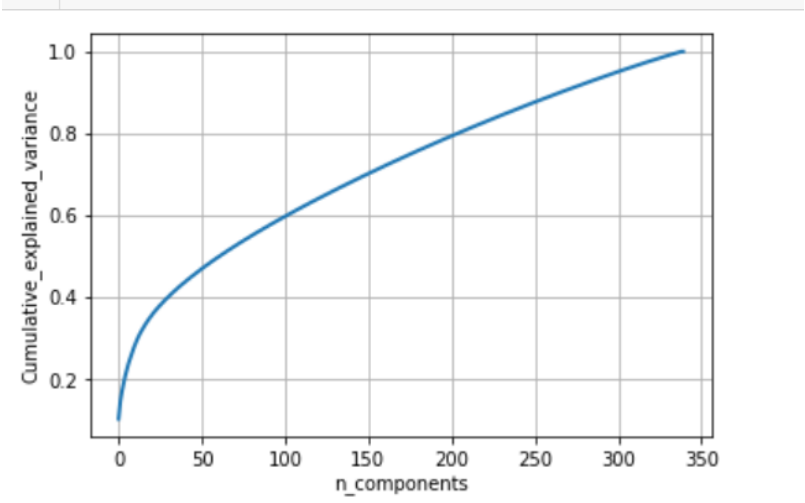
Solution: Both data set has high dimensional data.
Below are few of the statistic we observed:
Data Imbalance:
For some descriptor we observed some minor imbalance exists which need further up-sampling

Descriptor	Label 0	Label 1	Minor Imbalance
C0: 1	97	33	Yes
C0: 2	77	53	
C0: 3	83	257	Yes
C0: 4	51	289	Yes
C0: 5	146	194	
C0: 6	200	140	

Scope of Feature selection:
This data set is high dimensional.So applied dimension reduction technique (PCA) to understand potential option remove unwanted features. After applying PCA dimention reduction on train1 data , we observed 120 features can capture all information



Similarly for PCA 350 feature can do the same job

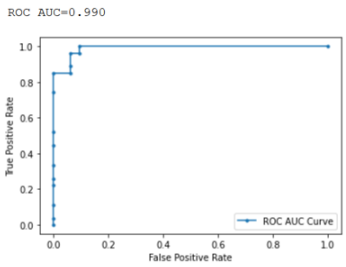


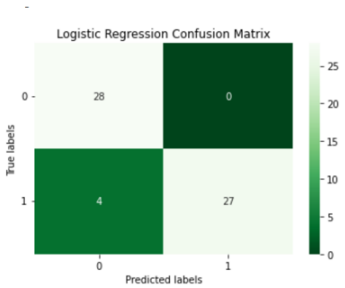
So there are some scope of reducing features.That can potentially improve performance.But final selection will be done only after applying various classifiers and analysing performance.

3. (points) [Brief introduction/motivation: Describe briefly the reason behind choosing a specific method for predicting each of the descriptors (could show plots or tables summarizing the scores of training different model)]

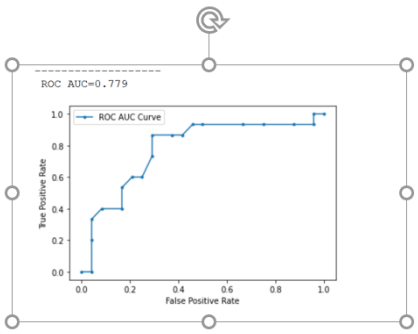
Solution:

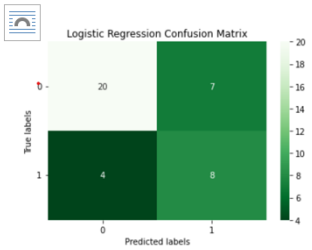
This data set contain huge dimensions , so it is challenging to visualise data graphically with few dimension, so we evaluate performance of various classifiers based metrics like f1 score,accuracy score,Confusion matrix,ROC Curve visualization etc.And select best suited classifier based on performance. Even for selected Classifiers we tried Parameter tinning using GridSearchCV to get best result.We also tested all the model with or without applying dimension reduction technique.Finally we feel accuracy is good (in Kaggle)with out any dimension reduction and it does not impact performance drastically.
C01, C02, C03, C04.C05:
We selected logistic regression with proper parameters tuning with GridsearchCV.We got good accuracy in parameters like accuracy score,confusion matrix and ROC curve.We also tried Bagging and Boosting ensemble on this data, but found only Logistic regression giving good accuracy.
Below is respective Confusion matrix and ROC curve
CO 1:Logistic regression



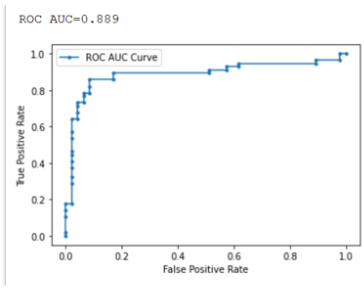


C02:Logistic regression

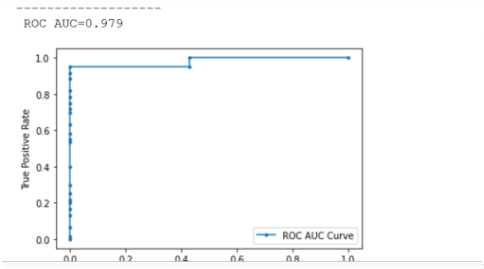


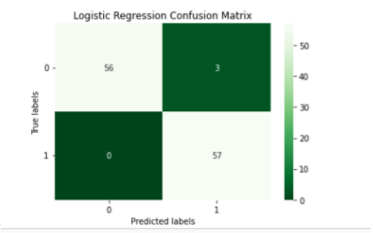


C03:Logistic regression

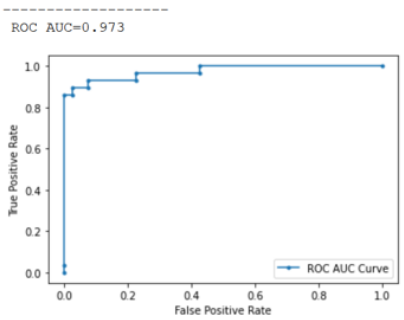


C04:Logistic regression



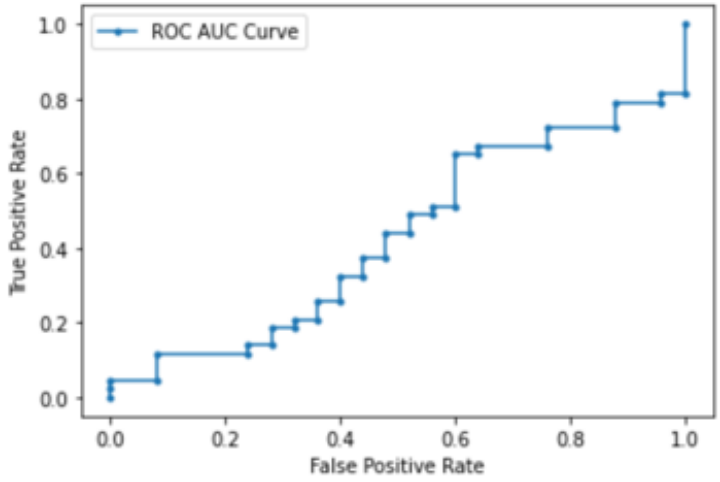


C05:Logistic regression



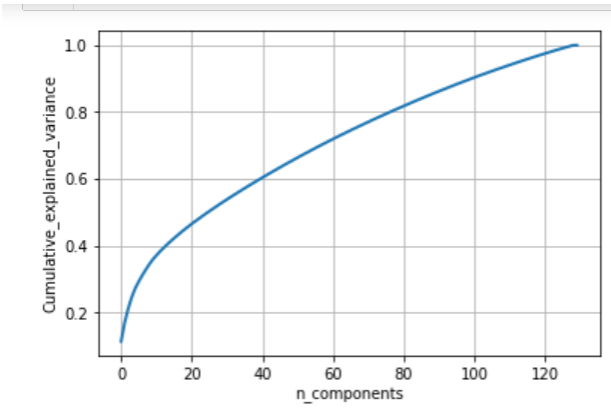
C06:ADABoost Claasifier

We found this model selection very challenging .As it is not giving very good accuracy with weak classifiers.So we opt for ADABoost classifier in order to get improved accuracy .

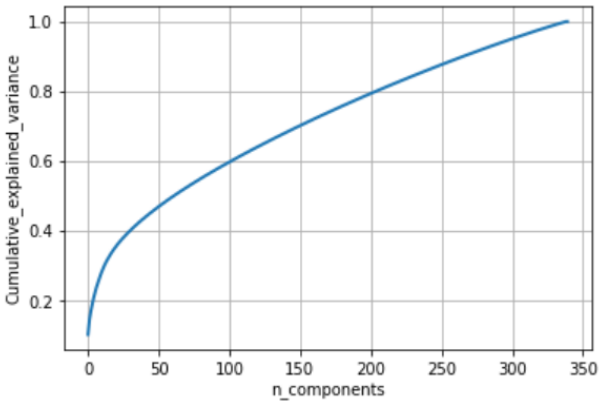


4. (points) [As the data comes from a clinical setting, model interpretation is also an important task along with prediction. Describe briefly your preferred choice of methods if you are asked to determine the significant genes behind the regulation of the endpoints or restrict the number of features/genes.]:

Solution: Our first preferred choice is doing Principal component analysis (PCA), which is a technique for reducing the dimensionality of high dimensional data sets and increasing interpretability but at the same time minimizing information loss.
After doing PCA on data set1 the results are as follows



100 features explain around 90 percent of the variance. From 22283 features to 90 features which is really helpful.
After doing PCA on data set2 the results are as follows



In data set2 250 features explain around 90 percent of the variance. From 54675 features to 250 whcih is really helpful.

After that we will think about scaling the data which standardize features by removing the mean and scaling to unit variance.

5. (points) [Share your thoughts on each of the endpoints: whether easy/difficult to predict]:

Solution:

CO: 1	moderate
CO: 2	moderate
CO: 3	moderate
CO: 4	easy
CO: 5	easy
CO: 6	difficult

6. (points) [Add any additional information: like challenges faced or some details that would help us to better understand the strategies that you have utilized for model development]:

Solution: As the data is imbalanced and has a very high number of features, it is not feasible to use simple classification algorithms like linear regression or naive Bayes. We must use algorithms that can withstand imbalanced data.
We got the good results only after doing Hyper parameter tuning. Logistic Regression uses class weights following the class distribution. Class-weights are the extent to which the algorithm is punished for any wrong prediction of that class. So it comes in handy to work with imbalance distribution. It is possible that even better performance can be achieved with non-default values of other hyper parameters of logistic regression.
We used ensemble methods which also gives good performance with imbalanced data.
After seeing the data, we tried many combinations of the models and finally decided that Logistic regression and Ada Boost classifiers are giving good results