Course Name: Statistical learning for big data
Course Code: MVE440

Devosmita Chatterjee
Personnummer: 910812-7748

June 12, 2020

Exercise 3.2: Classification and variable selection

1 Methods

1.1 Exploratory data analysis

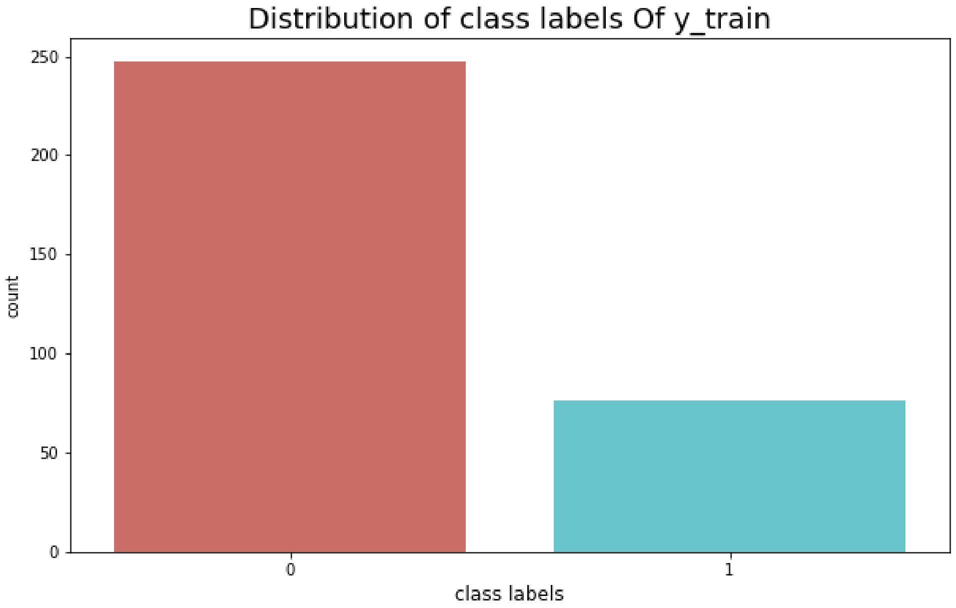
An exploratory data analysis is performed with data 'classification.npz'. Data size, data type, missing data, duplicate data are listed in table 1.

Table 1: *The table presents the exploratory data analysis with data 'classification.npz'.*

Exploratory Data Analysis						
Data	**Row size n**	**Column size p**	**Data type**	**Missing values**	**Duplicate rows**	**Duplicate columns**
X_train	323	800	float	0	0	0
X_valid	323	800	float	0	0	0
y_train	175	1	int	0	-	-
y_valid	175	1	int	0	-	-

The distribution, count and variability of the binary response variable 'y_train' is shown in figure 1. The classes are imbalanced since the ratio of 0's to 1's is 3.25.

Figure 1: *The figure shows the counts of observations in each class of y_train.*



1.2 Data standardization

Data standardization of a feature means to scale the observations of the feature with mean 0 and standard deviation 1 given by the formula

$$Z = \frac{X - mean(X)}{sd(X)}$$

The training data 'X_train' and the validation data 'X_valid' are standardized around mean 0 and with standard deviation 1.

1.3 Classification

From exploratory data analysis, we find that the training response variable 'y_train' is binary and the number of features is greater than the number of observations (p > n) in the dataset which motivates to choose the following two penalized logistic regression methods for classification and feature selection:-

1. L1–regulated logistic regression, and
2. Elastic net–regulated logistic regression.
The training data 'X_train' is analysed with the above two methods.

L1 regularization penalizes the absolute sum of coefficients (L1 penalty) - $\hat{\beta}^{Lasso} = \arg\min_{\beta \in R^p} |Y - X\beta|^2 + \lambda|\beta|$. L1 regularization enforces the β coefficients of one of the two highly correlated variables to be zero.

Elastic net regularization penalizes both the sum of squared coefficients and the sum of absolute coefficients (linear combination of L2 penalty and L1 penalty) - $\hat{\beta}^{Elastic} = \arg\min_{\beta \in R^p} ||Y - X\beta||^2 + \lambda(\frac{1-\alpha}{2}||\beta||^2 + \alpha|\beta|)$. Elastic net regularization enforces the β coefficients of one of the two highly correlated variables to be lower and other to be zero.

The hyperparameters studied are 1. inverse of regularization strength (C = $1/\lambda$) and 2. algorithm used to solve the optimization problem (solver). The best hyperparameters for the methods are obtained by exhaustively searching over a specified grid of parameter values for logistic regression estimator which are listed in table 2.

Table 2: *The table presents the methods' best hyperparameters using grid search algorithm.*

Methods' Best Hyperparameters		
Method	**Inverse of regularization strength**	**Algorithm to use in the optimization problem**
l1–regulated logistic regression	1	liblinear
Elastic net–regulated logistic regression	0.1	saga

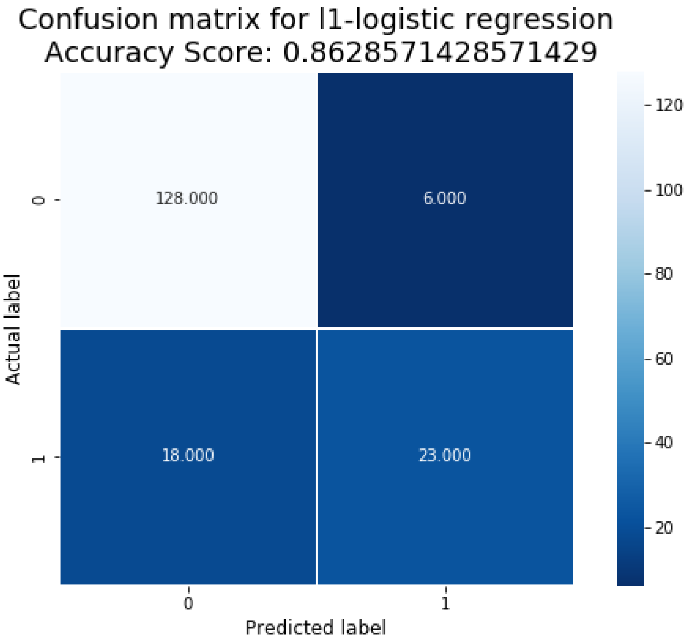
2 Results

2.1 L1-regulated logistic regression method performance

l1-regulated logistic regression method performance is studied on the validation data.

Accuracy score is the mean accuracy of validation class labels given by accuracy score = correct predictions/total predictions = correct predictions/(correct predictions + incorrect predictions). Figure 2 shows that there are 128+23=151 correct predictions and 6+18=24 incorrect predictions. The accuracy score of l1–regulated logistic regression with all predictors on the validation set is 0.862857148571429.

Figure 2: *The figure shows the confusion matrix for l1-logistic regression with all predictors on the validation set.*



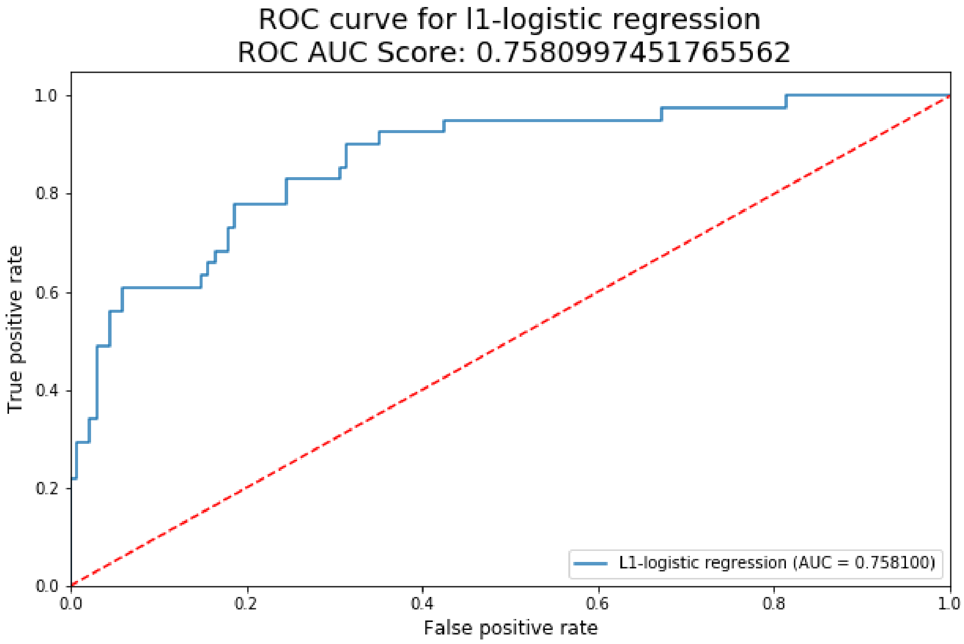
Classification report consists of precision, recall, f-beta score and support. Precision is used to evaluate the classifier's ability to not label a sample as positive if it is negative. Precision is the ratio of the number of true positives to the total number of false positives and true positives. Recall is used to evaluate the classifier's ability to find all the positive samples. Recall is the ratio of the number of true positives to the total number of false negatives and true positives. F-beta score is the weighted harmonic mean of precision and recall where beta = 1.0 implies that recall and precision are equally important. Support is the number of occurrences of each class in y_valid. The classification report of l1–regulated logistic regression with all predictors on the validation set is represented in table 3.

Table 3: *The table presents the classification report of l1–regulated logistic regression with all predictors on the validation set.*

	precision	**recall**	**f1-score**	**support**
0	0.88	0.96	0.91	134
1	0.79	0.56	0.66	41
accuracy			0.86	175
macro avg	0.83	0.76	0.79	175
weighted avg	0.86	0.86	0.85	175

Receiver Operating Characteristic (ROC) curve is used to evaluate the classifier's output quality. The ROC curve presents true positive rate on the y-axis and false positive rate on the x-axis. The red dotted line represents the ROC curve of a purely random classifier. A good classifier stays as far as possible from the red dotted line. Top left corner point is considered to be the ideal point and larger area under curve is usually better. The receiver operating characteristic (ROC) curve for l1-logistic regression with all predictors on the validation set is shown in figure 3. The ROC area under curve (AUC) score is 0.7580997451765562.

Figure 3: *The figure shows the ROC curve for l1-logistic regression.*



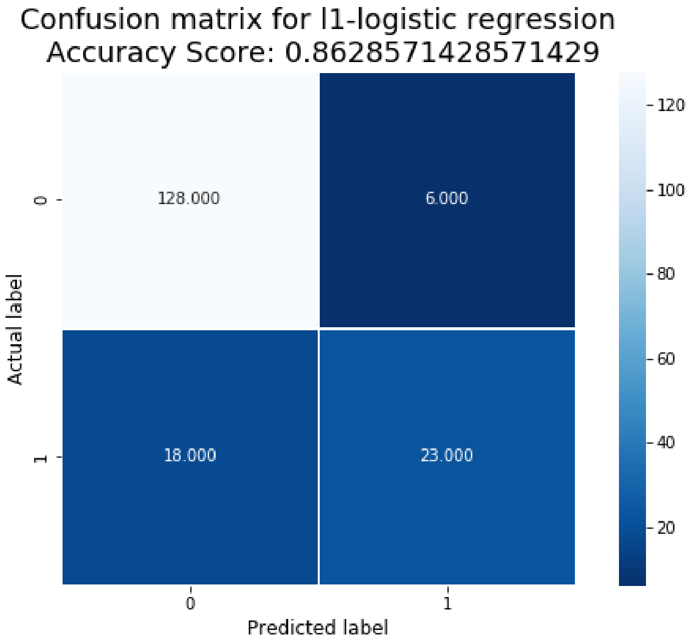
L1–regulated logistic regression selected 122 best predictors using feature selection property which is represented in table 4. After selecting the best predictors, the method performance is again evaluated on the validation set.

Table 4

Best predictors		
Total features	**Selected features**	**Features with coefficients shrank to zero**
800	122	678

Figure 4 shows that there are 128+23=151 correct predictions and 6+18=24 incorrect predictions. The accuracy score of l1–regulated logistic regression with best predictors on the validation set is 0.8628571428571429.

Figure 4: *The figure shows the confusion matrix for l1-logistic regression with best predictors on the validation set.*



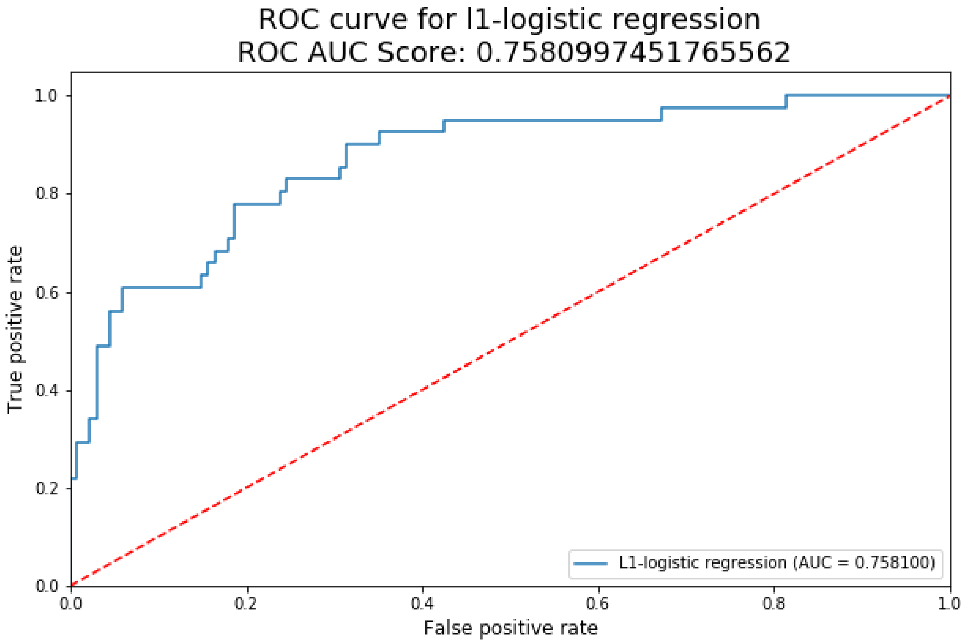
The classification report of l1–regulated logistic regression with best predictors on the validation set is represented in table 5.

Table 5: *The table presents the classification report of l1–regulated logistic regression with best predictors on the validation set.*

	precision	**recall**	**f1-score**	**support**
0	0.88	0.96	0.91	134
1	0.79	0.56	0.66	41
accuracy			0.86	175
macro avg	0.83	0.76	0.79	175
weighted avg	0.86	0.86	0.85	175

The ROC curve for l1-logistic regression with best predictors on the validation set is shown in figure 5. The ROC area under curve score is 0.7580997451765562.

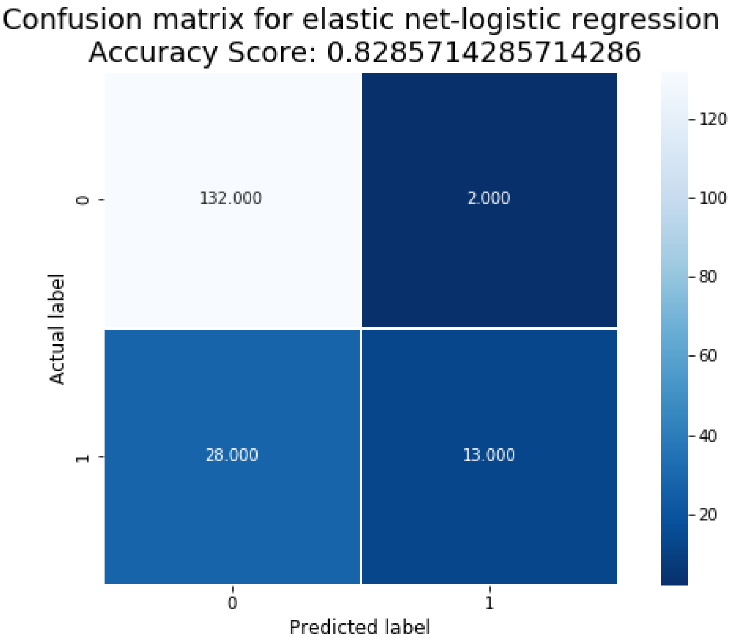
Figure 5: *The figure shows the ROC curve for l1-logistic regression.*



2.2 Elastic net-logistic regression method Performance with all predictors

In the section, elastic net-logistic regression method performance is studied on the validation data. Figure 6 shows that there are 132+13=145 correct predictions and 2+28=30 incorrect predictions. The accuracy score of elastic net–regulated logistic regression with all predictors on the validation set is represented in 0.8285714285714286.

Figure 6: *The figure shows the confusion matrix for elastic net-logistic regression with all predictors.*



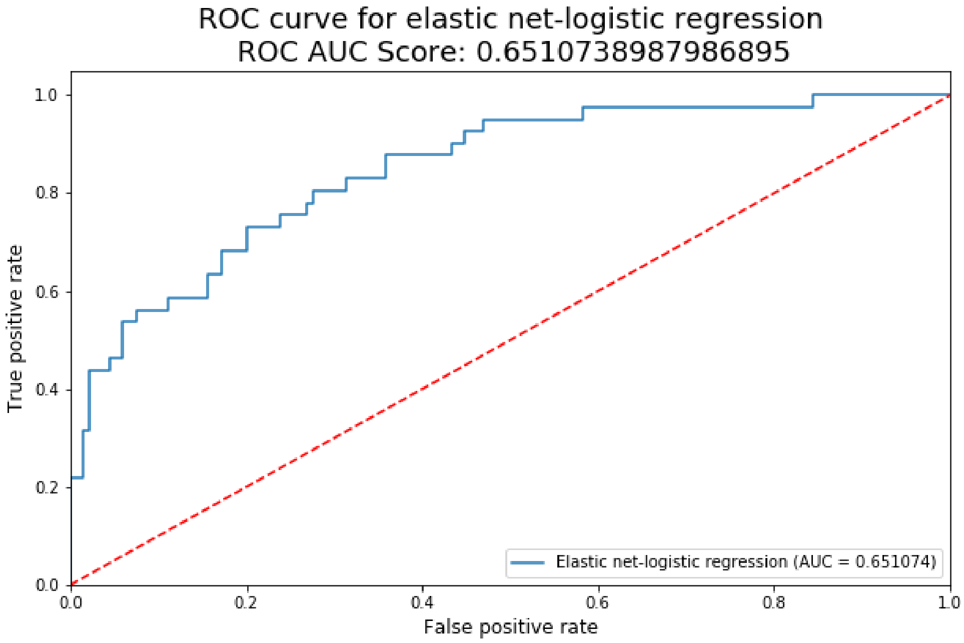
The classification report of elastic net–regulated logistic regression with all predictors on the validation set is represented in table 6.

Table 6: *The table presents the classification report of elastic net–regulated logistic regression with all predictors.*

	precision	recall	f1-score	support
0	0.88	0.96	0.91	134
1	0.79	0.56	0.66	41
accuracy			0.86	175
macro avg	0.83	0.76	0.79	175
weighted avg	0.86	0.86	0.85	175

The ROC curve for elastic net-logistic regression with all predictors on the validation set is shown in figure 7. The ROC area under curve score is 0.6510738987986895.

Figure 7: *The figure shows the ROC curve for elastic net-logistic regression.*



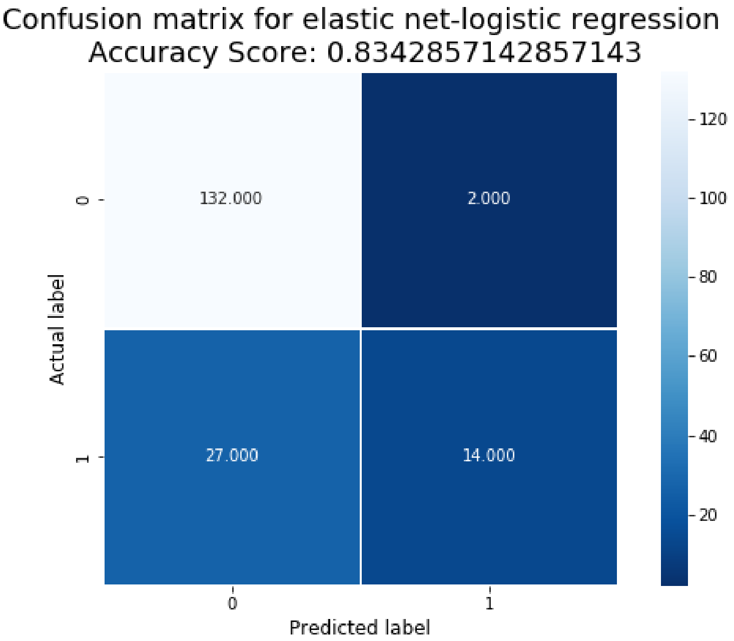
Elastic net–regulated logistic regression selected 59 best predictors which is represented in table 7. After selecting the best predictors, the method performance on the validation data is studied.

Table 7

Best predictors		
Total features	**Selected features**	**Features with coefficients shrank to zero**
800	59	732

Figure 8 shows that there are 132+14=146 correct predictions and 2+27=29 incorrect predictions. The accuracy score of elastic net–regulated logistic regression with best predictors on the validation set is 0.8342857142857143.

Figure 8: *The figure shows the confusion matrix for elastic net-logistic regression with all predictors.*



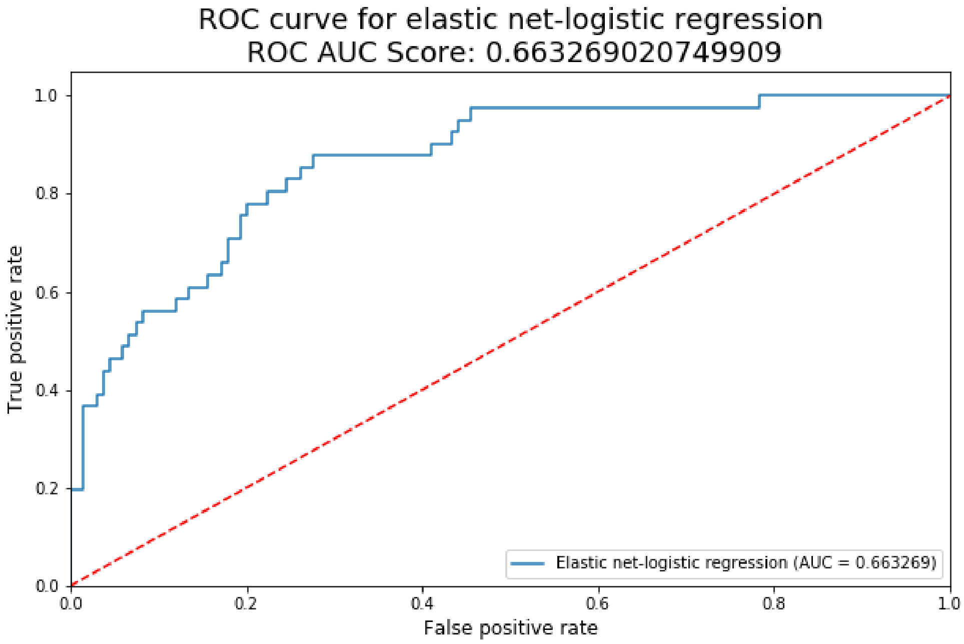
The classification report of elastic net–regulated logistic regression with best predictors on the validation set is represented in table 8.

Table 8: *The table presents the classification report of elastic net–regulated logistic regression with all predictors.*

	precision	**recall**	**f1-score**	**support**
0	0.83	0.99	0.90	134
1	0.88	0.34	0.49	41
accuracy			0.83	175
macro avg	0.85	0.66	0.70	175
weighted avg	0.84	0.83	0.81	175

The ROC curve for elastic net-logistic regression with all predictors on the validation set is shown in figure 9. The ROC area under curve (ROC AUC) score is 0.663269020749909.

Figure 9: *The figure shows the ROC curve for elastic net-logistic regression.*



3 Discussion

In the section, we discuss about the following results:-

1. L1 regularization does a sparse variable selection since it generally selects one highly correlated variable over the other. Elastic net regularization minimizes the importance of irrelevant features and do a sparse selection simultaneously. Therefore, the logistic regression used for the elastic net regularization remove more non-important predictors from the data than the logistic regression used for the l1 regularization.

2. The comparison between the accuracy scores of l1–regulated logistic regression with all predictors and best predictors on validation set is represented in table 9. With best 122 predictors, the accuracy score of l1–regulated logistic regression remains same as the accuracy score of l1–regulated logistic regression computed with all 800 predictors on the validation set.

Table 9: *The table presents the comparison between the accuracy scores of l1–regulated logistic regression with all predictors and best predictors on the validation set.*

Comparison Table		
Method		**Accuracy score**
L1–regulated logistic regression	**With all 800 predictors** **With best 122 predictors**	0.8628571428571429 0.8628571428571429

The comparison between the ROC area under curve (AUC) scores of l1–regulated logistic regression with all predictors and best predictors on the validation set is represented in table 10. With best 122 predictors,the ROC AUC of l1–regulated logistic regression remains same as the ROC AUC of l1–regulated logistic regression with all 800 predictors on the validation set.

Table 10: *The table presents the comparison between the roc auc scores of l1–regulated logistic regression with all predictors and best predictors on validation set.*

Comparison Table		
Method		**ROC AUC score**
L1–regulated logistic regression	**With all 800 predictors** **With best 122 predictors**	0.7580997451765562 0.7580997451765562

3. The comparison between the accuracy scores of elastic net–regulated logistic regression with all predictors and best predictors on validation set is represented in table 11. The accuracy score of elastic net–regulated logistic regression with

best 59 predictors on the validation set becomes slightly more than the accuracy score of elastic net–regulated logistic regression computed with all 800 predictors on the validation set.

Table 11: *The table presents the comparison between the accuracy scores of elastic net–regulated logistic regression with all predictors and best predictors on the validation set.*

Comparison Table		
Method		**Accuracy score**
Elastic net–regulated logistic regression	**With all 800 predictors**	0.8285714285714286
	With best 59 predictors	0.8342857142857143

The comparison between the ROC area under curve (AUC) scores of elastic net–regulated logistic regression with all predictors and best predictors on the validation set is represented in table 12. The ROC AUC score of elastic net–regulated logistic regression with best 59 predictors becomes slightly more than the ROC AUC score of elastic net–regulated logistic regression with all 800 predictors on the validation set.

Table 12: *The table presents the comparison between the roc auc scores of elastic net–regulated logistic regression with all predictors and best predictors on validation set.*

Comparison Table		
Method		**ROC AUC score**
Elastic net–regulated logistic regression	**With all 800 predictors**	0.6510738987986895
	With best 59 predictors	0.663269020749909

4. The accuracy of l1–regulated logistic regression method with 122 predictors remains same as the accuracy of l1–regulated logistic regression method with all 800 predictors. With around one seventh of the total number of predictors, the accuracy of l1–regulated logistic regression method remains constant.

The accuracy of elastic net–regulated logistic regression method with 59 predictors becomes slightly higher than the accuracy of elastic net–regulated logistic regression method with all 800 predictors. With around one fourteenth of the total number of predictors, the elastic net–regulated logistic regression method performs little better.

5. In terms of feature selection, elastic net–regulated logistic regression method performs better than l1–regulated logistic regression method. However, the overall accuracy of l1–regulated logistic regression method is better than elastic net–regulated logistic regression method.

1 Exercise 4.1: High-dimensional clustering

1 Methods

1.1 Exploratory data analysis

An exploratory data analysis is performed with data 'clustering.npz'. Data size, data type, missing data, duplicate data are listed in table 13.

Table 13: *The table presents the exploratory data analysis.*

Exploratory Data Analysis						
Data	**Row size n**	**Column size p**	**Data type**	**Missing values**	**Duplicate rows**	**Duplicate columns**
X	302	728	float	0	0	0

1.2 Data Normalization

Data normalization of a feature means to scale the observations of the feature between 0 and 1 given by the formula

$$Z = \frac{X - min(X)}{max(X) - min(X)}$$

The data 'X' is normalized between 0 and 1.

1.3 Dimensionality reduction

From exploratory data analysis, we find that the number of features is greater than the number of observations (p > n) in the dataset which motivates to choose tSNE for dimensionality reduction of the data 'X'. t-Distributed Stochastic Neighbor Embedding (tSNE) is an unsupervised, non linear technique used for visualizing high dimensional data. tSNE is a probabilistic approach.

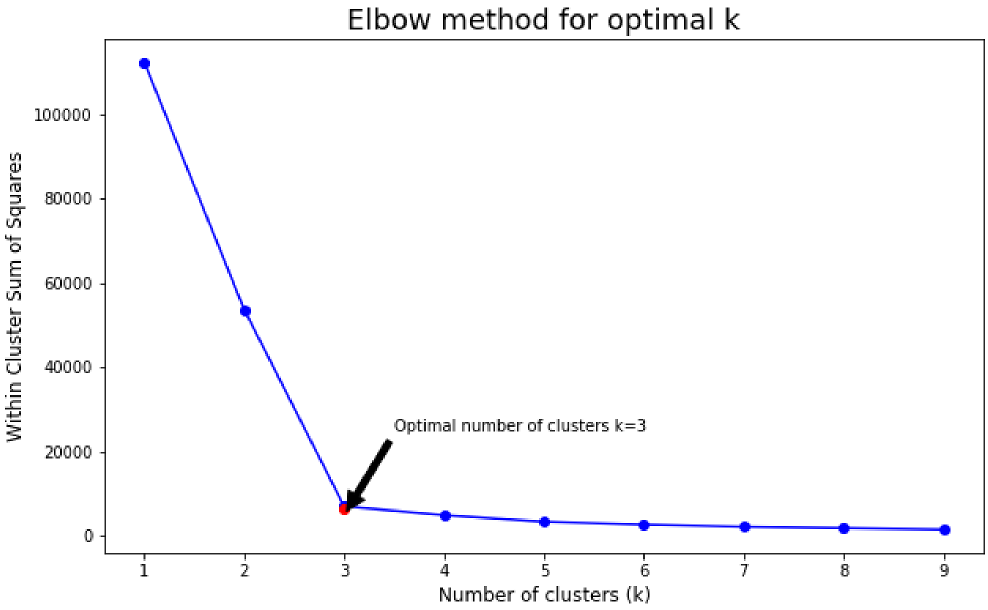
1.4 Clustering

For k-means clustering, Elbow method is used to find the optimal number of clusters in the dataset. Then the tSNE reduced data is colored according to the kmeans clustering with optimal number of clusters k.

2 Results

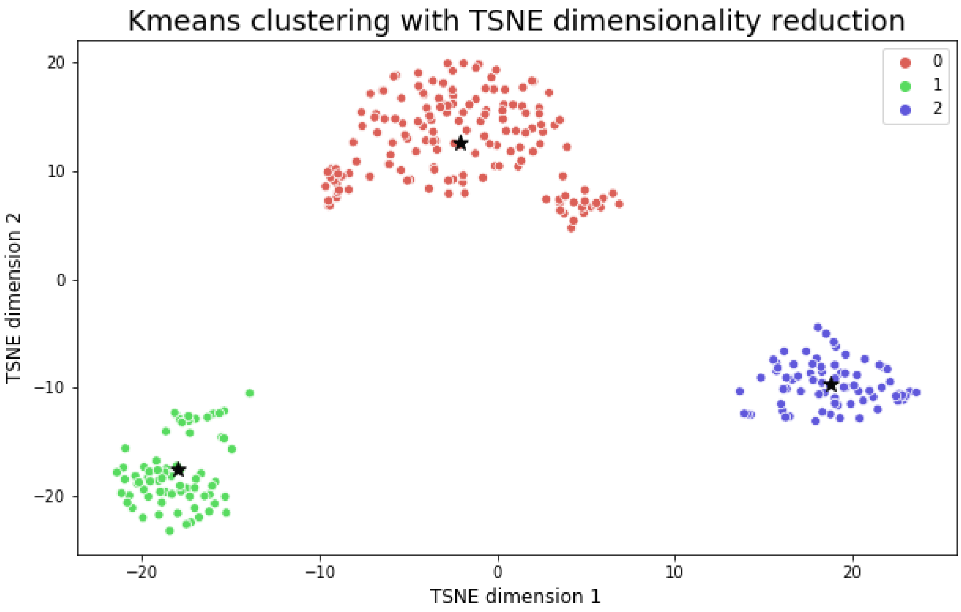
In Elbow method, to choose an optimal k for k-means clustering, the Within Cluster Sum of Squares (sum of the squared distance between each data point in the cluster and the cluster centroid) is plotted for different values of k. We will see in the plot the trend that the error decreases as the number of clusters increases. We select the optimal number of clusters where the change in Within Cluster Sum of Squares begins to slow down. The relationship between the number of clusters and the Within Cluster Sum of Squares is shown in figure 10. The optimal number of clusters is k=3.

Figure 10: *The figure shows the optimal number of clusters.*



To find clusters in the dataset, the tSNE reduced data is colored according to the kmeans clustering with optimal number of clusters k = 3. Kmeans clustering with tSNE reduced data is shown in figure 11. The cluster centroid for each cluster is marked with black star in figure 11.

Figure 11: *The figure shows the Kmeans clustering with TSNE reduced data.*



Again, the full data 'X' of order (302,728) is fitted to a kmeans model with number of clusters 3 and same random state as before. To find which variables are most indicative of cluster 1, we use the following formula

$$max(Dist(C1, C2) + Dist(C3, C1)) = max(|C1 - C2| + |C3 - C1|)$$

where $C1$, $C2$ and $C3$ are centers of cluster 1, cluster 2 and cluster 3 respectively. The five variables that are most indicative of cluster 1 are 319, 677, 137, 62 and 523 presented in table 14.

Table 14: *The table presents the best five variables that are most indicative of cluster 1.*

Best five descriptors					
Feature	**Center of cluster 1**	**Center of cluster 2**	**Center of cluster 3**	**Distance between centers of cluster 1 and cluster 2**	**Distance between centers of cluster 3 and cluster 1**
319	-0.110547	0.043650	0.154197	0.256835	0.411032
677	-0.079864	0.280001	0.359864	0.047954	0.407818
137	0.093998	-0.071044	0.165042	0.180406	0.345448
62	-0.087592	-0.078165	0.009427	0.312855	0.322283
523	-0.083907	0.132395	0.216302	0.104091	0.320393

To find which variables are most indicative of cluster 2, we use the following formula

$$max(Dist(C1, C2) + Dist(C2, C3)) = max(|C1 - C2| + |C2 - C3|)$$

where $C1$, $C2$ and $C3$ are centers of cluster 1, cluster 2 and cluster 3 respectively. The five variables that are most indicative of cluster 1 are 677, 403, 152, 545 and 523 presented in table 15.

Table 15: *The table presents the best five variables that are most indicative of cluster 2.*

Best five descriptors					
Feature	**Center of cluster 1**	**Center of cluster 2**	**Center of cluster 3**	**Distance between centers of cluster 1 and cluster 2**	**Distance between centers of cluster 2 and cluster 3**
677	-0.079864	0.280001	0.359864	0.047954	0.767683
403	-0.063805	0.177788	0.241594	0.011071	0.472116
152	-0.060826	0.153556	0.214382	0.013379	0.415385
545	0.037939	-0.132944	0.170883	0.027294	0.369059
523	-0.083907	0.132395	0.216302	0.104091	0.328513

To find which variables are most indicative of cluster 3, we use the following formula

$$max(Dist(C2, C3) + Dist(C3, C1)) = max(|C2 - C3| + |C3 - C1|)$$

where $C1$, $C2$ and $C3$ are centers of cluster 1, cluster 2 and cluster 3 respectively. The five variables that are most indicative of cluster 1 are 62, 416, 493, 677 and 86 presented in table 16.

Table 16: *The table presents the best five variables that are most indicative of cluster 3.*

Best five descriptors					
Feature	**Center of cluster 1**	**Center of cluster 2**	**Center of cluster 3**	**Distance between centers of cluster 2 and cluster 3**	**Distance between centers of cluster 3 and cluster 1**
62	-0.087592	-0.078165	0.312855	0.312855	0.616283
416	-0.076302	-0.077930	0.283738	0.283738	0.569105
493	-0.063753	-0.066675	0.244263	0.244263	0.491449
677	-0.079864	0.280001	0.047954	0.047954	0.455772
86	0.060841	0.043903	0.208899	0.208899	0.400859

3 Discussion

In the section, we discuss about the following results:-

1. TSNE is a powerful dimensionality reduction method. TSNE reduces the dimension of the dataset from p = 728 to p = 2.

2. When the tSNE reduced dataset is colored according to the kmeans clustering, we find three distinct clusters in the dataset. TSNE can handle well separated clusters as well as class imbalances.

3. TSNE yields visually pleasing results. In order to represent high dimension data on low dimension, nonlinear manifold, it is necessary that similar data points must be close together, which tSNE does.

4. The most indicative variables that are responsible for a specific cluster formation are found by computing the distance of the specific cluster center from the other two cluster centers for the variables. So when this distance parameter for a particular variable is high, then the variable has a large contribution for the specific cluster formation and therefore, most indicative of the specific cluster. The five variables that are most indicative of cluster 1 are 319, 677, 137, 62 and 523. The five variables that are most indicative of cluster 2 are 677, 403, 152, 545 and 523. The five variables that are most indicative of cluster 3 are 62, 416, 493, 677 and 86.