IE 582 Fall 20 Final Project

Ceren Özkan
Emre Kara
Şiar Bozyer

Istanbul, Turkey

1. Introduction

Classification is an important sub-field of machine learning discipline. There are two types of supervised learning. One is regression in which values of a quantitative dependent variable are predicted and the other is classification where the predicted variable is of class labels (qualitative values). When there the classification problem has a two labeled dependent variable, the problem is called binary classification. In this study a binary classification problem is investigated. [1]

The data analyzed in this study has 60 independent variables named as $\{x_i, i = 1, 2, ..., 60\}$ and a single dependent variable named y. The variables have no meaningful real-life names and no knowledge is available concerning the domain of the data. Therefore, during the study, no prior knowledge is used as a guidance in the data analysis and because of this, the effort was spent on developing models that would have as much generalization power as possible.

One important feature in this study is that the data has an imbalanced nature in terms of the classes of the dependent variable. However, it is important to predict observations from the minority class as well even more important in some data sets. Therefore, the second aim of the study is to build a model which would be able to catch instances coming from the minority class as accurate as instances coming from the majority class.

1.1. Proposed Approach

For best model based on AUC and BER scores, different classification models were tested (such as Decision Trees (DT), Random Forest (RF) and Stochastic Gradient Boosting (SGB)) on raw data. Then less important and less variance variables were determined and excluded in datasets. Then determined models (RF) were applied to new dataset. 10 fold cross validation with repetition, variable reduction and oversampling are used in the study. Details were explained in "Approach" section.

1.2. Descriptive Analysis

On train dataset, "y" column has only 2 values, "a" or "b". And its distribution is:

a	**b**
1565	509

Table 1: Table caption

0.7545 of data belongs to class "a". Therefore, the records in test dataset belong most likely to class "a".

In train dataset, 50th and 52nd columns have only "0" value. They cannot show any variance. They can be excluded.

Class	**Number of 0 Points**
a	2074
b	2074

Table 2: Table caption

Columns in datasets mostly have binary values. For x26, x37, x46, x49, x57, x59 columns have few records of "1", less then 25. They create small variances, therefor its effect for class selection can be excluded.

Class	**0**	**1**
x26	2057	17
x37	2073	1
x46	2059	15
x49	2053	21
x57	2072	2
x59	2061	13

Table 3: Table caption

x1, x8, and x14 columns distribute normally. X9, x10 and x11 have continuous values. X5, x6 and x7 distribute uniformly, and x27 and x32 have chi-squared distributed data. These columns can be scaled with appropriate to their distributions.

For x36 and x42, their records consist of mostly "0" values. For nonzero records, they can be varied. Therefore, different approaches can be applied on columns. They can be excluded dataset for creating unregular variances, or It can be created dummy variables which has binary values (0 – zero values, 1 – nonzero values), or they scaled with and without 0 values.

For X5, x6 and x7, they have values from 0 to 18 which distributed uniformly. Columns can be thought as "Categorical". One way to search these columns are to create dummy variables and assign each category to these variables. Then correlated categories can be discarded. Histograms for the relevant columns can be found in Appendix A.

Modeling raw dataset and finding importance is another way to preprocess. Model determined before can be compiled, and then column importance values can be determined. Less important columns can be excluded and compile model with new dataset.

2. Literature Survey

Imbalanced classification of the data requires some manipulation. In literature, resampling is proposed to overcome imbalanced classification problems. Thus, literature review is performed on the description of resampling, methods, and application. There are mainly 3 types of resampling: oversampling, undersampling and SMOTE. Oversampling increases the number of minority class data points in the training set by randomly replicating existing minority class members. [2] Random undersampling works in a similar way, however, decreases the number of majority class data points. Briefly, both methods try to balance the classes. SMOTE method allows the classifier to build larger decision regions that contain nearby instances from the minority class. [3]

Performance measurement is crucial in construction of a good model. As a final measurement, the average of AUC (area under curve) and BER (balanced error rate) is used to evaluate the performance of the model. AUC measures the ability of the model to avoid errors during classification using the receiver operating characteristic (ROC) curve. The AUC is equivalent to the probability that the classifier will rank a randomly chosen agreement higher than a randomly chosen error. [4] BER is the average of the rate of errors in each class. Also, most of the models gives accuracy, sensitivity, and specificity rates. Comparing those arguments also gives a good sense of performance.

To control the performance measures and gain the best results, 10-fold cross-validation technique is used. By basic definition, it consists in the controlled or uncontrolled division of the data sample into two subsamples, the choice of a statistical predictor, including any necessary estimation, on one subsample and then the assessment of its performance by measuring its predictions against the other subsample. [5] Thus, using cross-validation in model evaluation results in better results, since the accuracy control is assessed within the model.

3. Approach

Aim of the project is to perform a classification on a given dataset. Dataset has 2 classes in the output vector, and traditional classification methods are applicable. Thus, as a base, Lasso, Decision Trees (DT), Random Forest (RF) and Stochastic Gradient Boosting (SGB) methods are tried. Firstly, these methods are tried on raw data, without reduction or manipulation, to get a model that can fit our model best. This approach resulted in selecting Random Forests as the best model since they gave the highest accuracy results.

In the control function of the training models, 10-fold cross-validation is used. Cross-validation creates subsamples within the train data and compare the accuracies within. Thus, instead of manually manipulating the columns and try the model, it is performed automatically.

After selecting a best model, to improve the results, data should be manipulated. The first way to perform is the reduction of the instances. To choose the columns that will be extracted, mainly 2 methods are used: variance and importance of the columns. It is observed that there were highly imbalanced instances in the train dataset, and high imbalance results in very low variance. Thus, columns with variances close (or equal) to zero are deducted. For example, columns x52 and x54 were consisting of all 0 points. This results in zero variance, thus, no impact on our model and hence, they are extracted. Since best model for the raw data is random forest, results on random forest is analyzed and columns are ordered in importance. Last few columns with lowest importance are deducted from our data to gain better results-and expected- better results are achieved.

Train set of this project is consisting of 60 features and a class vector, which involves 2 class: a and b. There is a total of 1565 variables in class a and 509 variables in class b. It can be seen that there is almost

a 1:3 ratio between classes, which can be considered as a "class imbalance problem". There exist many solutions to this problem, and the best and most used one is applying resampling methods. 3 methods are suggested as a resampling method in previous section, and each method is tried in evaluation. As a result, oversampling gives the best solution for classification of our dataset. This is because of the following nature of the methods. SMOTE method is mostly applicable k-NN, however, our data is not a good fit to perform k-NN classification. Thus, data manipulation with SMOTE was not a good application. Undersampling is slightly worse because minority class is consisting of 509 data points-which is not sufficient to train a good model. Hence, oversampling method is a good way to manipulate the data.

As another approach, standardization is performed on the continuous data. To decrease variances of the instances, continuous data points are reduced between the interval [0,1]. Also, some columns are introduced with value-or-not transformation. It is observed that, some of the instances have a high number of zeros, and some random numbers. Those numbers are not meaningful separately, hence, binary transformation (with the condition that if a value is different than 0, transform it to 1) is applied. However, both these methods gave worse results. This is due to the fact that by applying much transformation, the nature of the data is corrupted. Thus, even though the data is improved, this does not always generate better results.

4. Results

To achieve the best results, importance detection and instance reduction due to importance is crucial. Also, random forest algorithm is the best model to fit the data. After applying random forest to raw data, it is concluded that columns x1, x2, x4, x7, x10, x11, x19, x26, x29, x33, x47, x49, x50, x52 and x57 does not affect the model significantly. Thus, for the following models, they are extracted from the data. These columns are detected with the importance object in the model. Also, looking at the data, some of them have significantly low variance, thus the results are expected.

Besides applying reduction methods, resampling methods affect the test performance significantly. After trying the main three methods for resampling, best results are achieved with oversampling. Basically, oversampling reduces the imbalanced attribute of output data and results in a better model.

As described in the "Approach" section, other applications such as standardization of continuous data, value-or-not transformation and different classification approaches are tried on the train data. However, none of them gave better results. This concludes that manipulating data is not always results in a better fit.

To sum up, in the best results are achieved with column reduction, oversampling and random forest model. In the test set, AUC metric is calculated as 0.9256 and BER is calculated as 0.8293. As the final result, a total score of 0.8774 is achieved, which in general is a good result in predictive models.

5. Future Work

For future works,

• New classification models (which may be found in literature),

• Clustering before classification,

• Different preprocessing approaches (penalized column addiction, different scaling approaches)

• New sampling methods (which may be found in literature),

can be applied on datasets and compare new results from best scored model based on AUC, BER etc. results.

6. Conclusion

In classification exercises, there exists many factors that should be controlled. Firstly, a good fit for the model should be determined. However, only finding a good fit is not enough when there exists bias, imbalanced classes etc. Thus, data manipulation should also have a significant role in the predictive models.

Data manipulation can include many approaches. To solve imbalance problems, resampling approaches are highly recommended. This allows artificial data generation or random data reduction, to achieve a balanced classifier. To eliminate unnecessary columns, importance measures should be considered. After training the raw data with some model, from the related object, importance and the usage of the columns can be determined. For the following models, unimportant columns can be extracted, and so, total variance would be decreased. This results in a better fit for the data. However, extracting excessive features results in worse models, since under fitting is introduced and necessary data can not be extracted from the data.

When training a model, performance measures should always be controlled. To have better results, cross-validation method is proposed in this project. This method allows subsampling and having the better fit to the model. After predicting the test data, accuracy measures such as sensitivity, specificity, area under curve,

balanced error rate etc. should be checked and controlled if there is an improvement in the model. Performance measures allows to comment the results better and selecting the best predictive model.

To sum up, there are many criteria that should be checked in predictive models. One model or one approach does not always work in general, it highly depends on the data. Thus, data manipulation is also crucial. However, to achieve best results, one should think of every possible application in data, and by training and testing, best fit is possible to achieve.

Appendix A. Appendix

Project Codes: https://bu-ie-582.github.io/fall20-karaee/files/Project/Project.html

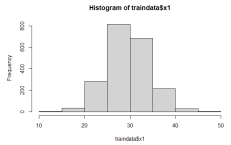

Figure A.1: Histogram of x1

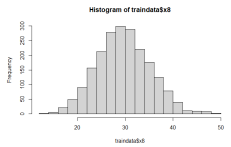

Figure A.2: Histogram of x8

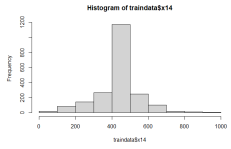

Figure A.3: Histogram of x14

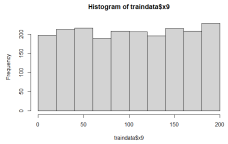

Figure A.4: Histogram of x9

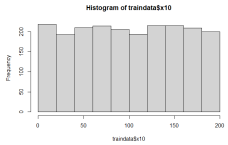

Figure A.5: Histogram of x10

References

[1] T. H. R. T. Gareth James, Daniela Witten, An Introduction to Statistical Learning with Applications in R, Springer, Reading, Massachusetts, 2013.

[2] A. Liu, J. Ghosh, M. C., Generative oversampling for mining imbalanced datasets, Generative Oversampling for Mining Imbalanced Datasets (2007).

[3] V. Garcia, J. S. Sanchez, M. R. A., Exploring the performance of resampling strategies for the class imbalance problem, Trends in Applied Intelligent Systems (2010) 541–549.

[4] L. Cuadros-Rodríguez, P.-C. E., C. Ruiz-Samblás, Quality performance metrics in multivariate classification methods for qualitative analysis, Trends in Analytical Chemistry (2016) 612–624.

[5] M. Stone, Cross-validatory choice and assessment of statistical predictions, Journal of the Royal Statistical Society. Series B (Methodological 36 (1974).

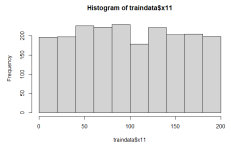

Figure A.6: Histogram of x11

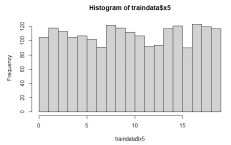

Figure A.7: Histogram of x5

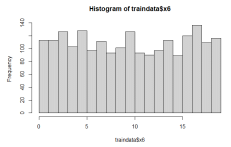

Figure A.8: Histogram of x6

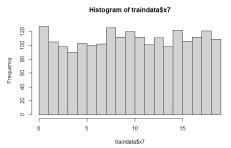

Figure A.9: Histogram of x7

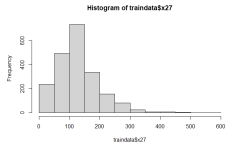

Figure A.10: Histogram of x27

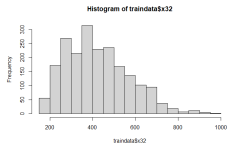

Figure A.11: Histogram of x32

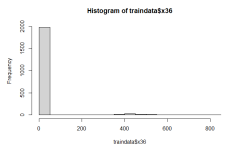

Figure A.12: Histogram of x36

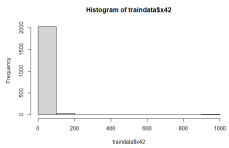

Figure A.13: Histogram of x42