STA160 Midterm Project Report: Analysis of Classification and Regression for Multivariate Dataset

Siyuan Li
University of California, Davis
sssli@ucdavis.edu

Abstract: Within the report, we will apply data analysis for two different multivariate datasets. The variables contain continuous and categorical data. The report will extend the exploration of various clustering and classification methodologies under two circumstances.

DATA ANALYSIS PART I: WHEAT SEED KERNEL:

I. INTRODUCTION

The first part of the project is concentrated on the seed.txt data. The resource is provided from Institute of Mathematics and Computer Science, and Department of Automatic Control and Information Technology at Cracow University of Technology. The data collected information of geometrical properties of kernels belonging to three different varieties of wheat, which is obtained by a soft X-ray technique. Specifically, the data is constructed with seven geometric parameters related to wheat kernels, and all of parameters are continuous real-valued variables.

II. DESCRIPTIVE DATA EXPLORATION

Following the first step of data analysis, we carried on an exploration of basic test statistics of the dataset. The summary result is listed in below table with corresponding attribute information.

	Mean	Median	Min	Max	Std
1. area A	14.8475	14.3550	10.5900	21.1800	2.9097
2. perimeter P	14.5593	14.3200	12.4100	17.2500	1.3060
3. compactness C	0.8710	0.8735	0.8081	0.9183	0.0237
4. length of kernel	5.6285	5.5235	4.8990	6.6750	0.4431
5. width of kernel	3.2586	3.2370	2.6300	4.0330	0.3777
6. asymmetry coefficient	3.7002	3.5990	0.7651	8.4560	1.5036
7. length of kernel groove	5.4081	5.2230	4.5190	6.5500	0.4915

Table 1

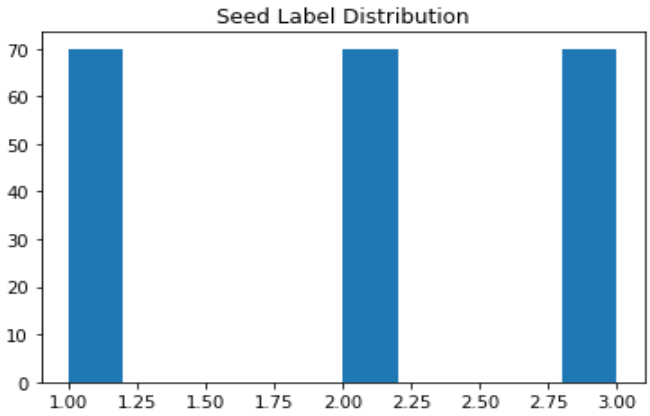

Figure 1

To apply further classification techniques on the seed dataset, we check the distribution of data labels for having better random sample selection. Based on the test statistics and the previous information, we know that labels are categorized into three group. For convenience of mathematical calculation, we named three labels in numerical form with "1", "2" and "3". Additionally, the given data resource demonstrates that three kinds of wheat type have the same number of examples. In the next level, we will apply various multiclass classification techniques to predict the seeds label, explore their model accuracy and discuss the best features for appropriate model selections.

III. Classification Analysis

The first step to carry on classification process is to split the original data points into two categories with training dataset and testing dataset. According to the previous analysis, our randomization process can be chosen with a test size equal to 0.3. The reason is because the label distribution is strictly even presented. Notably, there is no missing value in the dataset, we will perform the random process directly.

In the report, we explored four different multiclass classification methods for the seed dataset. Those techniques are Logistic Regression, Decision Tree Regression, Random Forest and Adaboost. The analysis contained a calculation of average accuracy scores based on 30 tests for each model. The exploration also looked into a specific sample of the seed dataset, and applied above methodologies to find their confusion matrix and error matrix.

Additionally, we will also apply the Gridsearch function to explore the best parameters combination of the models such as Random Forest and Adaboost. The summary test statistics is listed in the below tables.

Multiclass Classification Methods	Scores
Logistic Regression	0.8963
Decision Tree Regression	0.8698
Random Forest	0.8719
Adaboost	0.7113

Table 2 Accuracy Scores Based on 30 tests:

According to the above accuracy table, we can have a direct comparison of model efficiency of all listed methodologies. Logistic regression model showed the best result as we are leading 30 sample tests for label prediction. Meanwhile, Adaboost method demonstrated an obvious shortage among the four techniques. It obtained the least accuracy score and the difference between the other three score value is nearly 0.15. However, the above result could not be directly used as a final conclusion to decide the Logistic Regression as the best model fit because we can see that the accuracy scores for Decision Tree Regression and Random Forest are close to its value. In the next level, we explored one determined sample data, and applied those classification techniques. We are hoping to compare the actual difference, and to search the parameters factors within each model.

The specific random sample data is predetermined by one fixed "random_state" number. Then we plotted the confusion matrix and error matrix for each four method individually. The plots are presented as follow

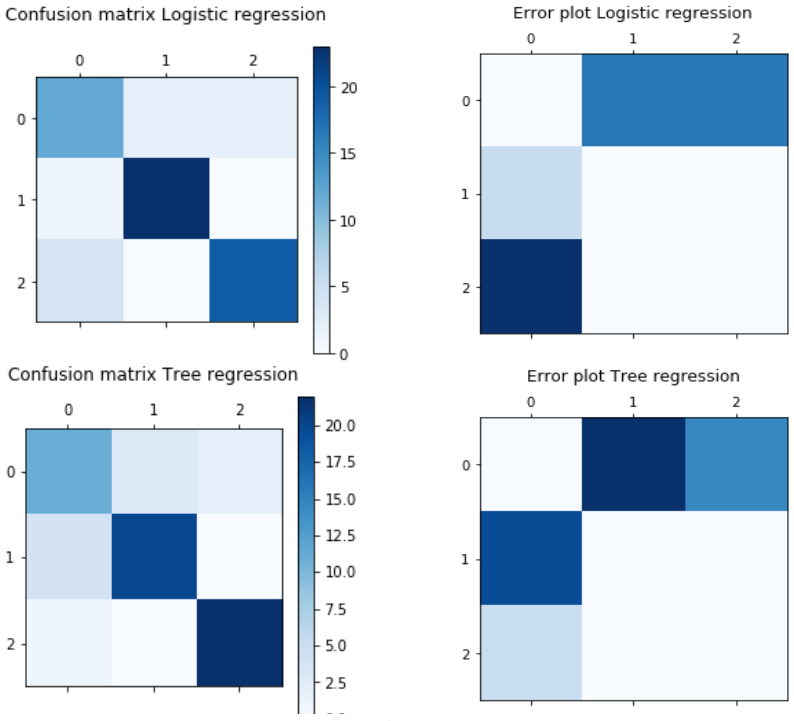

Figure 2

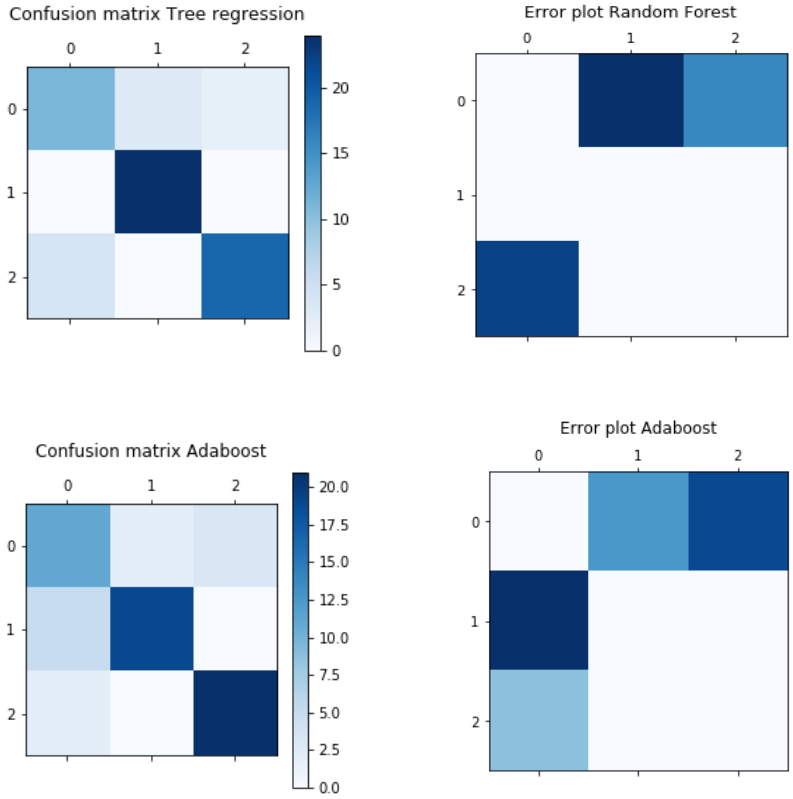

Figure 3

The above plots can provide a good estimation of how each classification model perform in prediction. By comparing the confusion matrix of each method, we find that they present a similar distribution. One particular difference will be Adaboost method, which be observed to have more prediction error for category "1" given the original label is "0". We may correspond the discovery with previous statistics result as Adaboost achieved the least accuracy score.

Then, we can research for the error matrix. We obtain them by obliterating the diagonal value of confusion matrix into zeros. We can find that the Logistic Regression model does not avhe the best prediction result. The graphs show us that prediction errors appearing for Random Forest method and Tree Decision Regression are distributed similarly as Logistic Regression. However, the difference can be further explored by the Gridsearch function.

For model parameter optimization, we perform Gridsearch function on the above two methods: Random Forest and Decision Tree Regression. We modified our gridsearch range for Random Forest by "n_estimators" and "max_features"; and same process for Decision Tree Rergession model by "max_depth". We calculate the best accuracy score and presented its parameters combination. The new result of four classification methods is summarized as below.

Multiclass Classification Methods	Scores	Parameter
Logistic Regression	0.8571	none
Decision Tree Regression	0.8843	“max_depth” = 7
Random Forest	0.9115	“n_estimators” = 50 “max_features” = 2
Adaboost	0.8095	none

Table 3 Based on one specific sample

The above two test statistics for Decision tree regression and Random Forest are obtained by Gridsearch function. As we approach the optimized parameters index for the two classification techniques, we can obviously observe that the accuracy scores for both two methods have been increased. For this predetermined sample data, we now can conclude the Random Forest prediction method will have the most efficient prediction result. For further reflection, we set out previous 30 random sample data test with the above optimized parameter index, and we achieved the following test statistics.

Multiclass Classification Methods	Scores
Logistic Regression	0.8963
Decision Tree Regression	0.8671
Random Forest	0.8783
Adaboost	0.7113

Table 4 Based on 30 sample tests

When we set the index back to random sample testing procedure, we find that the Random Forest method is not providing the most efficient prediction result given 30 independent model trials. In general, the Logistic Regression have a better performance regardless of any parameter’s modification. In general, considering the calculation cost and time consumption, Logistics regression is preferred for the seed dataset. Meanwhile, the possibility is existing as we might able to find an optimized parameters combination for the classification, which requires more time to research.

DATA ANALYSIS PART II: AUTOMOBILE

I. INTRODUCTION

The second part of the project is concentrated on the automobile data. The resource is provided from 1985 Ward's Automotive Yearbook. The data collected information of three important types of entities: the specification of an auto in terms of various characteristics, auto's assigned insurance risk rating, and its normalized losses in use as compared to other cars. Specifically, the data is constructed with 26 parameters related to auto's characteristics and detailed information, and parameters are consisting of categorical and continuous data.

II. DESCRIPTIVE DATA EXPLORATION

Following the first step of data analysis, we carried on an exploration of basic test statistics of the dataset. The summary result is listed in below table with corresponding attribute information. We find that there are 10 categorical variables and 15 continuous variables.

Categorical Parameter	Continuous Parameter
1. make 2. fuel-type 3. aspiration 4. num-of-doors 5. body-style 6. drive-wheels 7. engine-location 8. engine-type 9. num-of-cylinders 10. fuel-system	1. normalized-losses 2. wheel-base 3. length 4. width 5. height 6. curb-weight 7. engine-size 8. bore 9. stroke 10. compression-ratio 11. horsepower 12. peak-rpm 13. city-mpg 14. highway-mpg 15. price

Table 5

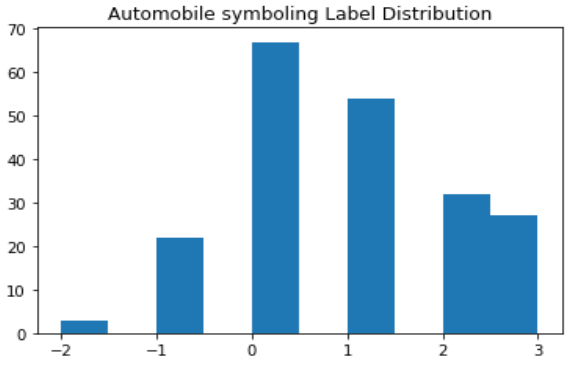

Figure 4

To apply further classification techniques, we check the distribution of data labels for having better random sample selection. (The label is set be the first column of data, Symbolling Index) we observe that the samples collected in the data source is not perfectly even distributed. As a result, when we apply the classification models the situation can cause bias and inaccuracy.

0	**1**	**2**	**3**	**-1**	**-2**
67	54	32	27	22	3

Table 6

Additionally, we plotted the histograms for categorical data within the automobile dataset.

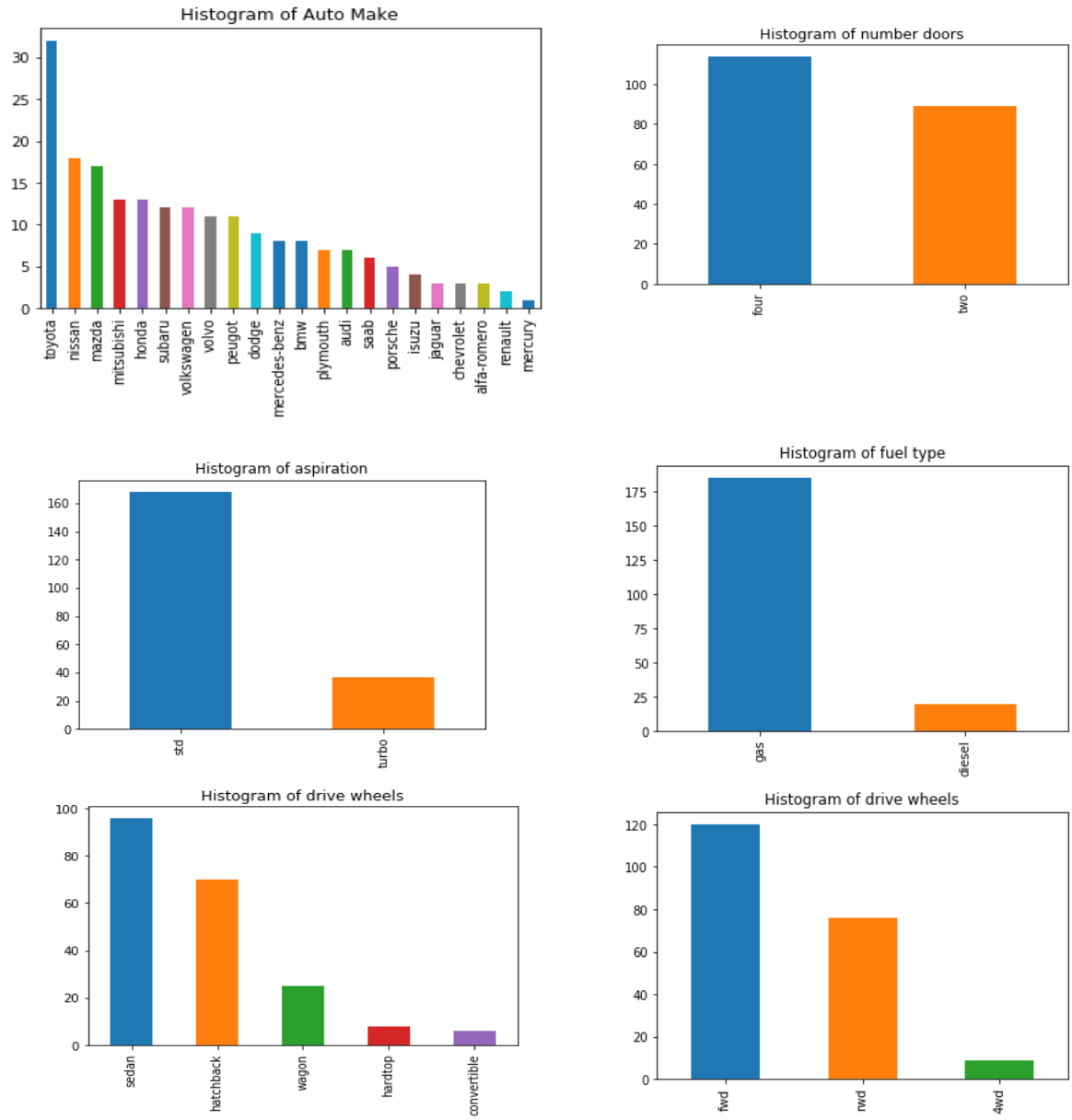

Figure 5

Next step after the basic descriptive statistical analysis of automobile dataset, we are hoping to apply classification techniques to predict the target label. In our analysis, we define the target label as automobiles' symbolling index. According to the introduction, we know the range of the index is from -3 to +3. The larger value infer that the auto is riskier for insurance. We treated the labels as a categorical data; therefore, we have seven classes. Notably, in the automobile dataset, only 6 class data points are contained.

However, as we can observe the data contains NaN value, we clean those data points and the final validate data resource consists on 159 autos' information. Here, we started to apply the different classification methods.

III. Classification Analysis

The automobile dataset contains 10 categorical variables, to perform Logistic Regression, Random Forest, and Decision Tree regression model. We need to factorize those data into numerical category for model fit. Based on the previous classification analysis on seed dataset, we have already observed that the Adaboost method doesn't show a great prediction result. As in in the second part of analysis, we will concentrate on the above first three techniques. Similarly, we calculated the accuracy score for the three methods. We are hoping to find their confusion matrix and error matrix based on one specific sample data. The test statistics are summarized as below.

Multiclass Classification Methods	Scores
Logistic Regression	0.6187
Decision Tree Regression	0.7520
Random Forest	0.7819

Table 7 Based on 30 random sample tests

With the above calculation result, we find that Logistic regression only have an accuracy rate as 61.87%, which is significantly less than the other two method. We observe that the Random Forest's accuracy rate is approximately 78.19%, which is a little bit higher than Decision Tree Regression model fit. However, the difference is not much large that lead us to decide which method is better for making prediction of the automobile dataset. Then we researched on one specific sample and would also utilize random search function to help us find the optimized parameters' index. The graphs of confusion matrix and error matrix are listed as follow.

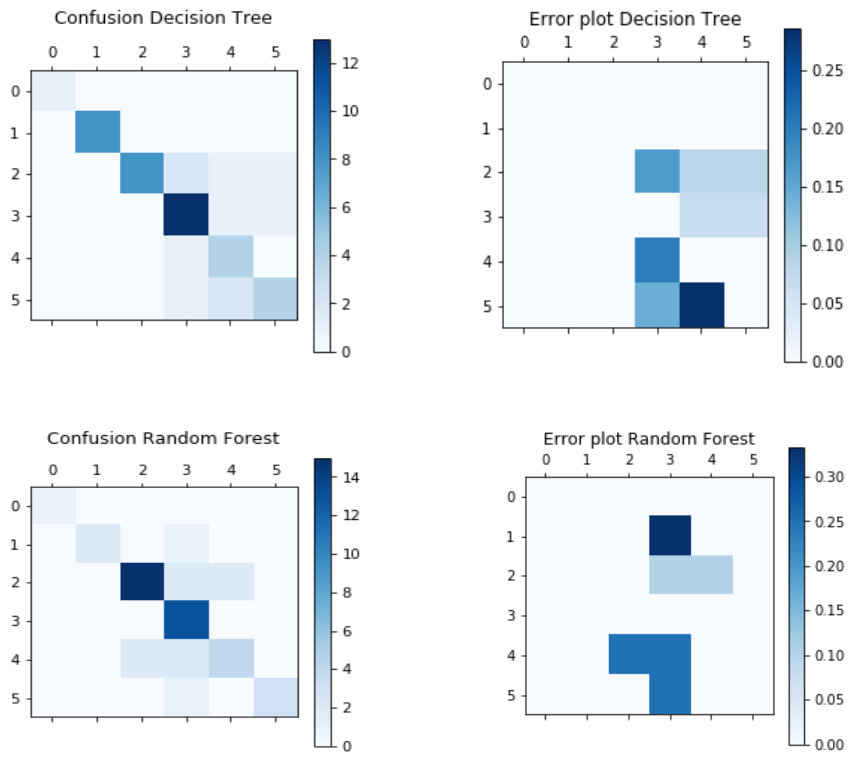

Figure 6

By comparing the confusion matrix and error matrix of Random Forest and Decision Tree regression, we can discover that under the method of Decision tree regression, model would produce more kinds of prediction error. Meanwhile, the Random Forest obtained an excellent level of prediction accuracy for category "0", "1", "4" and "5 category. It also has smaller error distribution area in its error matrix compared to the Decision Tree model. This advancement in graph support the previous basic calculations of accuracy scores with 30 random sample model tests. Therefore, we reasonable to conclude the Random Forest is the most efficient classification method for the automobile datasets.

In the next level, we hope to find the optimized parameters within Random Forest model for our prediction. The approach is achieved through "random.search" function. We summarized the final result as below.

	Accuracy scores	Parameter
Specific random sample	0.7207	"n_estimators" = 50 "max_features" = 3
30 random samples	0.8131	"n_estimators" = 50 "max_features" = 3
30 random samples	0.7819	none

Table 8

We find that by finding the optimized parameter index of "n_estimators" and "max_features", our accuracy rate for prediction will be improved. Random forest can be the best model fit.

CONCLUSION

We carried on classification methodology research on two different datasets, which finally lead us to conclude two model fit for each resource. The major approaches we selected are Logistic regression, Decision Tree Regression, Adaboost and Random Forest. Since both two datasets is in multivariate form and their target labels are not binary. We hope to choose the most efficient classification techniques that suitable for the multiclass case. There is certainly existing minor bias in our calculation as our random training sample dataset is still small. Meanwhile, we can observe that the distribution of label in seed data is even, while quite messy in automobile dataset. Thus, our training sample might fail to capture all labels data points, which also causing bias. To summary, seed data is consisting of all continuous variables, we find the most appropriate method is Logistics regression; automobile dataset contains both categorical and continuous variable, which is a more complicated condition. The best model fit we select is Random Forest. We hope to continue research the parameter index, and other combination of those above techniques to achieve a better prediction result later.

REFERENCES

[1] Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

[2] Oliphant, T. E. (2006). *A guide to NumPy* (Vol. 1). Trelgol Publishing USA.

[3] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

[4] McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).