Implementation of Outliers Identification and Handling Techniques for Classification

Aditya Tammeneni
Electrical Engineering Department
IIT Madras, Chennai, India
ee14b061@ee.iitm.ac.in

Gaurav Raina
Electrical Engineering Department
IIT Madras, Chennai, India
gaurav@ee.iitm.ac.in

Dhanajaya B L
Solverminds, Chennai, India
djay@solverminds.com

***Abstract*— In this paper, we provide a sequential and well-defined approach for identifying and handling outliers for classification problems. We implement these techniques on three datasets- i) Adult dataset, ii) Mobile price classification and iii) EEG eye state. The paper details in exploration, preparing, modeling and handling outliers. We consider three widely used machine learning algorithms for our predictions - i) KNN, ii) Logistic Regression and iii) Random Forest. We compare the performance of the models using accuracies, f1 score, and chi-square depending on the application. No unique outliers handling technique works well on all datasets and hence we need to choose depending on the dataset. Using these techniques, we observe an increase in accuracy of the models by 1%-2% for the Adult dataset and mobile price classification and no significant improvement for EEG eye state dataset.**

I. INTRODUCTION

Outliers are a very well studied topic in literature and many efficient methods have been proposed [2], [4]. We have implemented few of those methods to check the performance of these methods on real-world datasets. Some of the methods work well in general and others are dependent on data. We study the proposed methods for the classification problem. The classification has many real-world applications and widely used in fields like medical, financial, physical sciences, etc. The performance is analyzed based on three widely used machine learning algorithms, KNN, logistic regression & random forest and on three datasets - Adult dataset, mobile price classification dataset & EEG eye state dataset. The datasets are taken from the UCI Machine Learning Respository [3]. The algorithms used are easy to implement and works well in practice. We observe that the performance of these algorithms cannot be generalized and is data dependent. The aim of this implementation is not only to improve the accuracy of the model but also to improve the goodness-of-fit of the models.

The Adult dataset is about classifying the individual income range based on features like age, education, marital status, etc. In the EEG eye state dataset, the state of eye i.e whether it is closed or open is predicted based on readings from different EEG sensors placed on the head of an individual. Both these are binary classification problems. To check the performance on a multi-level classification, we used mobile price prediction dataset in which the price range is predicted using the various features of a mobile phone like ram, battery power, number of cores, etc. This is a 4 level classification problem.

TABLE I: Dimensions of Datasets

Dataset	Adult Data	Mobile Prices Classification	EEG Eye State
Instances	32561	2000	14980
Variables	15	21	15

The rest of the paper is organized as follows: Section II describes the three datasets used and the outliers identification and handling techniques are discussed in Section III. Section IV describes the algorithms used and their performance analysis. We conclude our work in Section V.

II. Problem Setting

A. Data Description

The dimensions of the three datasets can be found in Table I.

1) Adult Dataset: The Adult dataset has 32561 instances and each instance has a target variable income and 14 features like individuals age, education, occupation, marital status, etc. The objective is, given these features we need to classify the income of an individual as $\leq$ 50K or $>$ 50K. This is a binary classification problem dependent on 14 variables out of which 6 are numerical(age, fnlwgt, capital_gain, capital_loss, education_num and hours_per_week) and 8 are categorical(workclass, education, marital_status, occupations, relationship, race, sex, and native_country). All the categorical variables are nominal. The two levels of income $\leq$ 50K and $>$ 50K have 24720 and 7841 instances respectively. As can be seen, the dataset is slightly skewed with 24% of data belonging to one class and 76% to other.

The dataset has 7% missing values in terms of a number of rows with missing values and they are in workclass, occupation, and native country columns. The workclass has 5.6% missing values while the occupations and native country have 5.6% and 1.8% respectively.

2) Mobile Price Classification: The Mobile price classification dataset has 2000 instances, each corresponding to a different model of mobile phone. Each instance has

21 different variables corresponding to different features of the mobile phones like battery power, clock speed, mobile weight, internal memory, RAM, whether it has bluetooth or not, whether it supports 4G or not, etc. There are features related to physical aspects of the mobile(depth, screen height, screen width & mobile depth), camera and screen(front camera pixels, primary camera pixels, pixel resolution height, pixel resolution width, touch screen), memory(ram & internal memory), hardware related features(battery power, number of cores, dual sim, clock speed & talk time) & communication related features(3G, 4G, bluetooth & wifi). The task is to classify the range to which their prices belong. The target variable, price_range is divided into four levels 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost). Out of the 2000 instances, there are 500 instances belonging to each one of the four classes of price_range. The dataset has numeric and boolean variables and no categorical variables.

3) EEG Eye State: This dataset is obtained as a result of the experiment conducted by the authors of [6]. The dataset has 14980 instances each with 15 attributes out of which 14 are the readings of the sensors placed at different positions of the brain. The sensors AF3, F3, F7, FC5, T7, P7 & O1 are placed on the left half and AF4, F4, F8, FC6, T8, P8 & O2 are placed on the right half of the brain. The exact positions of these sensors can be found in [6]. In the experiment, a person is asked to randomly alter the state of his eyes and the readings were recorded. The experiment duration was for about 117 seconds. There are 8257 instances of class 0 which indicates an open eye and 6723 instances of class 1 which indicates a closed eye. The results of various classification algorithms are stated in [6] and [8].

B. Visualization

1) Adult Dataset: The data has 21790 samples of male and 10771 belonging to female of which 30% of the male population has income more than 50K whereas it is 10% of the female population. In terms of education, 87% of the people spent at least 9 years on education and after the schooling, most of them have gone to some colleges. Most of the people work in private sector and work for 40 hours a week.

The following groups of people tend to have more percentage of people with income > 50K:

- Age group of 40 to 55
- Positive capital gain or capital loss
- Number of education years ≥ 9
- Working hours per week of 40 to 60
- People who are married and staying together
- Asian Pacific Islanders and American whites

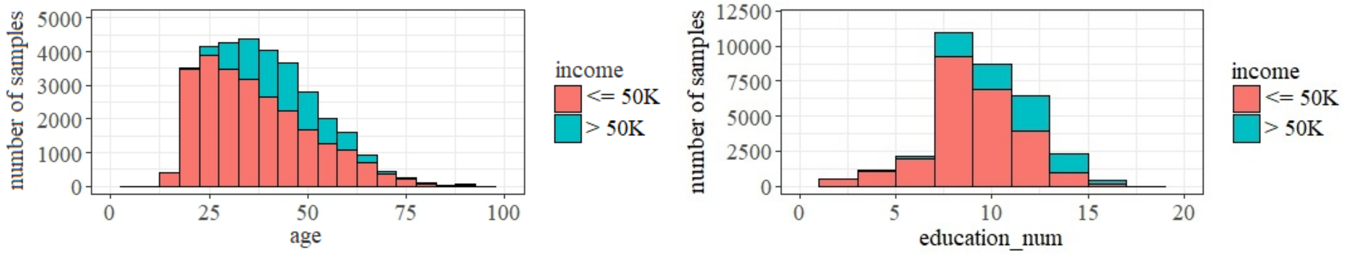
2) Mobile Price Classification: Few plots between some features are shown in Figure 2. Figure 2a, shows that ram and price range have a high correlation and it is 0.917. The price range has significant correlations only with ram, battery power, pixel resolution height(px_height) and pixel resolution width(px_width). Also, the observations are almost uniformly distributed in the ram and internal memory dimensions. Out of the 2000 instances mentioned, 1523 supports 3G and 1043 supports 4G. It is also interesting to note that there are 510 phones which support 4G but don't have a touchscreen. The primary camera megapixels are always more than the front camera megapixels. A new variable mobile_vol is created which is the volume of the mobile obtained from the height, width and depth and is compared to the weight of the mobile. It is observed that there no density value which is dominant among the mobiles. The talk time which is the duration for which the mobile can be used continuously without charging is independent of the battery power.

3) EEG Eye State: The mean values of the sensors which are either on top of the head or which are on the right half of the brain are more likely to be higher and it is lower for the ones which are towards the sides or at the back when the eye is closed. It can be observed from Figure 3b that, the features can be divided into two groups such that variables within the same group having strong correlations. All the features have a very less variance and their density plots are similar looking to that of the one in Figure 3a.

C. Statistical Analysis and Cleaning Data

The Adult dataset has missing values in workclass, occupations and native country. The filling has to be done in such a way that it preserves the initial distribution of the data. Few common methods for filling missing values- (i) Filling with mean and mode for continuous and categorical variables respectively. (ii) Filling based on the distribution of that variable and the target variable. (iii) Filling the required variable using its distribution with the variables which are correlated to it. The number of classes in some variables is high and we found that combining few classes improved the performance. For the variable native country, there are 42 classes(countries), we merged them based on their continents. The classes in variable education were merged based on the total number of years to achieve that degree. Similarly, the classes in the variables marital status, relationship, occupations, workclass and race were also merged based on the distribution and type of classes. We calculated the correlation between the variable and the target variable before and after merging these classes. The correlations have increased significantly for some variables as can be seen in Table II and thus increasing the accuracy and f1 score.

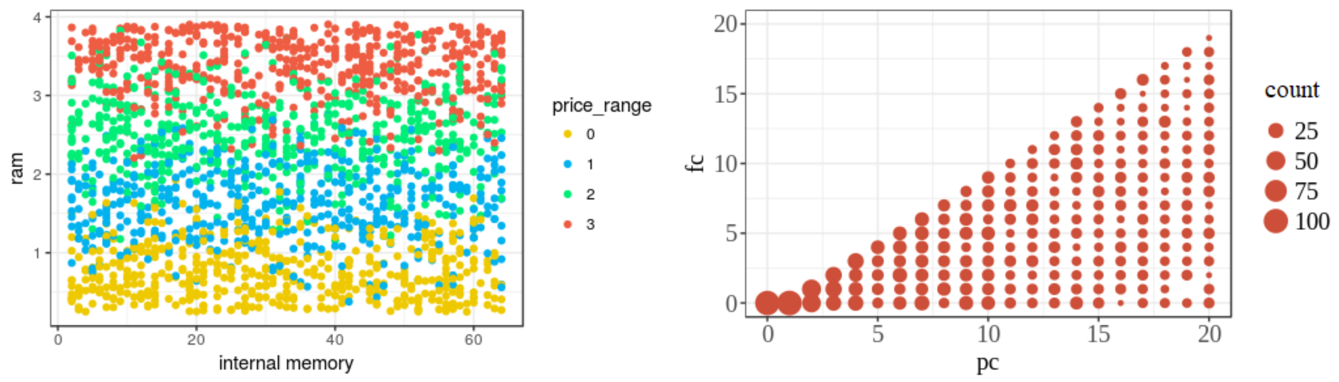
In the EEG eye state dataset, there are a few points which are around 10,000 times the interquartile range from the mean. So, these points are classified as outliers and removed from further analysis. Combining the variables which are similar based on the correlations in Figure 3b didn't cause any significant improvement in the performance of the model.



(a) Age with respect to income

(b) Education years with respect to income

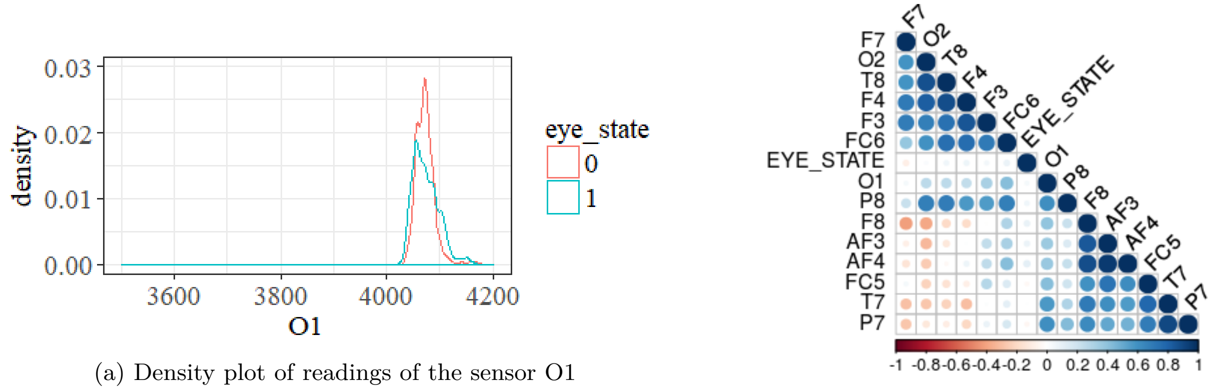
Fig. 1: Bar plots of variables with respect to income



(a) Ram vs internal memory with price range in GB

(b) Front camera(fc) vs primary camera(pc) in MP

Fig. 2: Relationships between variables of Mobile price classification dataset



(a) Density plot of readings of the sensor O1

(b) Correlation between variables

Fig. 3: Relationships between variables of EEG eye state dataset

TABLE II: Correlations with income before and after the variable transformation in Adult dataset

	capital_gain	capital_loss	workclass	education_num	marital_status	occupation	relationship	race	native_country
Before	0.306	0.206	0.003	0.460	0.564	0.047	0.527	0.121	0.031
After	0.423	0.20	0.211	0.477	0.611	0.477	0.534	0.137	0.10

III. Outliers Management

Outliers can significantly affect the model and they need to be handled in a well-defined approach. We follow the approach mentioned in [2].

A. Defining Outliers

1) Error Outliers: The type of outliers that lie at a distance from other data points due to inaccuracies in observations are called error outliers.

2) Interesting Outliers: Outliers that lie at a distance from rest of the data and are accurate data points and have potentially valuable information are called interesting outliers.

3) Influential Outliers: The data points that are accurate and alter the fit of the model and influence parameter estimates are called model fit and prediction outliers respectively. Both these come under influential outliers.

There are many other definitions for outliers in literature. 14 different definitions for outliers are mentioned in [2]. In this paper, we majorly focus on influential outliers. Though we discuss error outliers and interesting outliers, the work is majorly build towards handling the influential outliers.

B. Identifying Outliers

Identifying outliers is done using various single construct and multiple construct techniques. In single construct techniques, we use box plot and percentage analysis to identify potential error outliers. Using single construct techniques, we can detect the points which are away from the rest of the data only in one dimension and we need to do this separately for all the dimensions. Doing this we might miss out points which are near the mean in all the dimensions of the data but still lying away from the data. To be able to detect such points, multiple construct techniques like statistical methods, clustering [1], distance based approaches [5] are used. In multiple construct techniques, we use leverage values. The leverage values are the diagonal elements of the hat matrix. In case of multinomial classification with k classes, after performing multinomial logistic regression, we get k probability values each indicating the probability of the data point belonging to each one of those classes. So, in this case, we consider only the estimated probability of its actual class. The leverage values are calculated similarly to the binary classification problem using the corresponding error values.

Leverage values are a measure indicating the extent to which the data points are outliers. A high leverage value indicates that the data point is away from rest of the data points in the R^p space, where p is the number of features. A point which has a high value in any one of the dimension and around the mean in other dimensions need not necessarily have a high leverage value. A point which is away from the rest of the data is more likely to change the estimates of the parameters of the model which causes a change in the decision boundary which as a result are more likely to be influential outliers. The data points with a magnitude of leverage values higher than $2(p+1)/n$, where p is the number of predictors and n is the sample size are considered to be potential error outliers. From the threshold value, we identified 497 instances in the Adult dataset, 71 instances in Mobile price classification dataset and 1002 instances in EEG eye state dataset as potential outliers.

Potential error outliers should be cross verified against original records to declare them as error outliers. In our case, since the datasets are taken from online sources, it is not possible to cross verify them.But, in a situation where it is possible, it is recommended to so, resulting in the identification of error outliers.

Potential error outliers which are not error outliers become potential interesting outliers. We do not have any interest group to draw insights from, so there are no interesting outliers. The potential interesting outliers that are not interesting outliers are potential influential outliers.

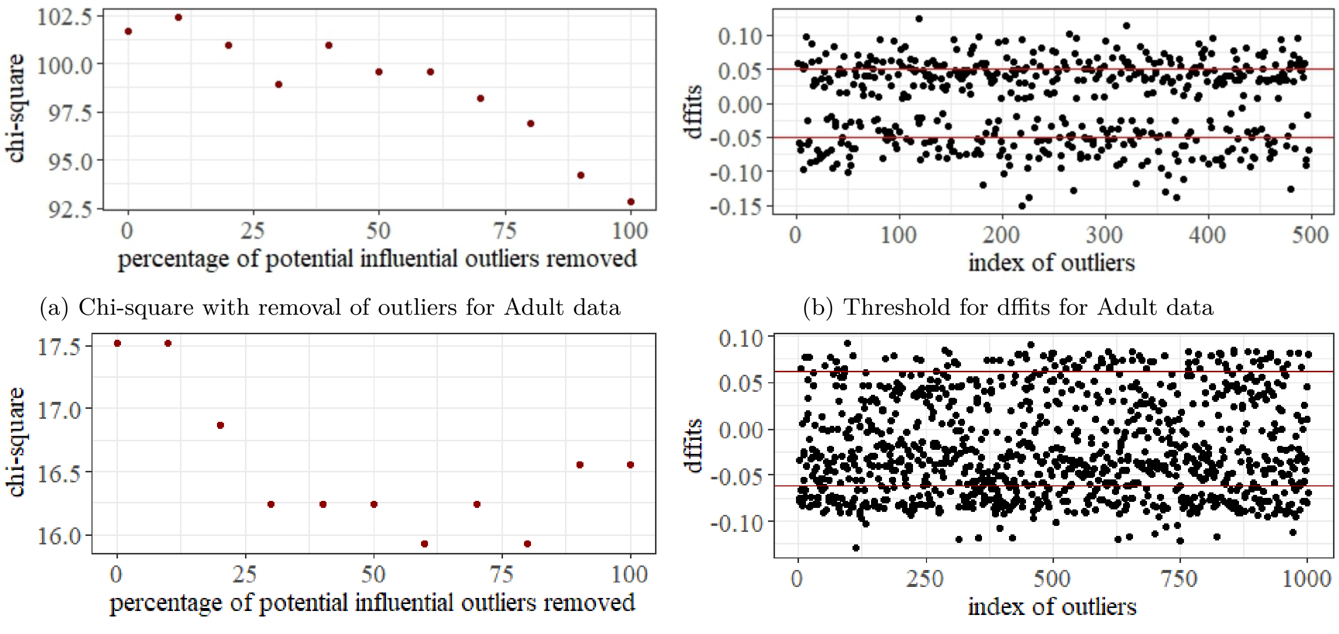
High leverage values doesn't mean that they influence the model in all cases. So, to identify model fit outliers, we observe the effect of each potential influential outliers on the model fit. The chi-square values are calculated for the models with and without these potential influential outliers and the difference is observed. If the chi-square value is significantly lower for the model without the data points, they can be classified as influential model fit outliers. But, this approach may not always help in identifying outliers in case of classification. In case of regression, observing the change in the output value by removal of a particular data point can be observed even if the change is in the order of 10^{-6}. But, this is not possible in case of classification, as the change might not be big enough to change the predicted class from 0 to 1 or vice-versa. So, we remove different groups of points and check if it is making any significant difference in the fit of the model. The potential outliers are removed incrementally based on their leverage values, and the chi-square values of the models are observed as shown in Figures 4a, 4c and 5. The chi-square values are decreasing with removing more points till some extent. This indicates that the model fit is better without those data points. It is also observed that by further removing the data points i.e the data points which are just under the threshold of leverage values in case of Adult data, the chi-square value starts to increase in case of Adult dataset. For EEG eye state, the chi-square value is minimum after removing 80% of the potential outliers. So, they are classified as influential model outliers. A downside of this approach of removing data points in groups is that few data points which are not actually outliers but are in the group with many other outliers might be classified as outliers.

For prediction outliers, we observe the effect of each potential influential outliers on parameter estimates. This is done by calculating Cooks distance, dffits which measures the difference in the estimated outputs of the models with and without the data point. Figures 4b and 4d shows the dffits values of the potential influential outliers. Threshold value is indicated by the red line for prediction outliers is $2\sqrt{(p+1)/n}$. The points whose magnitude is more than the threshold are classified as influential prediction outliers.

C. Handling Outliers

After applying the above identification techniques, we arrive at influential outliers and few ways to handle them are described below-

1) Keep them - We acknowledge the presence of the outliers and do not change anything related to the outliers. It is important to understand how well the



(a) Chi-square with removal of outliers for Adult data

(b) Threshold for dffits for Adult data

(c) Chi-square with removal of outliers for EEG eye state

(d) Threshold for dffits for EEG eye state

Fig. 4: Chi-square and dffits for Adult data and EEG eye state

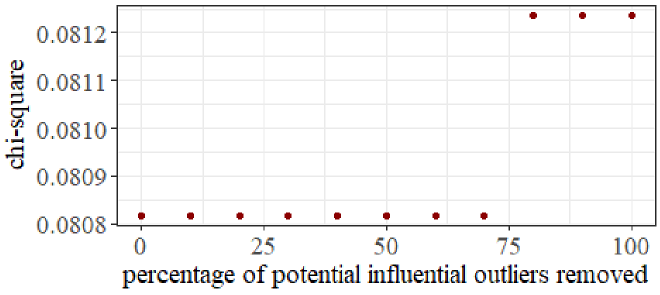


Fig. 5: Chi-square with removal of outliers for mobile price classificaiton

model fit is with the presence of the outliers to realize the effect of other handling techniques used. So, it is suggested to check the model fit with the presence of outliers.

2) Remove outliers - Outliers are not considered to build the models and for further analysis.
3) Remove wrongly classified outliers - The outliers are considered as the test data and a model is built using rest of the data. Using this model, the wrongly classified outlying points are removed from further analysis. Hence, a model is built without these points. In [7], a similar but more sophisticated method is implemented for the whole dataset and not just for the outliers on 53 datasets.
4) Non-parametric methods - Non-parametric methods [2] don't assume any underlying distribution of the data. So, we replace the variables which have very high values in them with their rank which reduces the influence of these points.

IV. Algorithms/Models

A. KNN

KNN is a non-parametric lazy learning algorithm. Non-parametric means that it does not make any assumptions on the underlying data distribution. This is useful in cases where the data does not follow standard distributions like Gaussian, etc. It is also called lazy because it does not use training data to make any generalizations. It uses all the training data to make any predictions. There is practically no training phase, but it has a costly testing phase as each test sample is compared with all the training samples. It calculates the Euclidean distance between the test sample and all the training samples and takes k nearest distance training samples or neighbors. A test sample is classified by the majority vote of its neighbors.

B. Logistic Regression

Logistic regression is a binary classification machine learning algorithm. It basically defines a Bernoulli distribution with parameter p which is dependent on the input independent variables. The aim of the machine learning algorithm is to find the parameter p in such a way that it defines the separation boundary between the two classes. Rather than choosing p that minimizes the sum of squared errors, logistic regression chooses parameters that maximize the likelihood of observations. To perform maximum likelihood estimate we take the log of the probability and take its derivative and equate it to zero. But this expression has no closed form solution so we use iterative non-linear optimization techniques like

TABLE III: Performance of various algorithms without removing outliers

	Adult Data	Mobile Prices Classification	EEG Eye State
KNN	84.9	91.5	97.8
Logistic regression	85.2	96.5	64.3
Random forest	84.4	74.5	69.9

TABLE IV: Performance of various handling techniques

	Adult Data	Mobile Prices Classification	EEG Eye State
Keeping outliers	84.9	91.5	97.8
Removing based on chi-square & dffits	85.8	93.7	97.3
Removing wrongly classified outliers	84.7	91.4	97.6
Non-parametric methods	85.1	92	95.2

Newton's method or gradient descent. These techniques provide a solution but it is a local minimum.

C. Random Forest

Bagging is an ensemble method in which multiple weak classifiers are trained with bootstrap samples to achieve a strong classifier. Random Forest is a bagging technique. If the correlation between the classifiers is high, then the overall performance tends to be low. When the decision trees are bagged, the nodes at the lower level(near to the root) tend to make similar decisions increasing the correlation between classifiers. In Random forest, to reduce the correlation between classifiers, a random set of variables of size t <p (typically t $=\sqrt{p}$) is considered at every split and the split is made only on one of these variables from this random set.

D. Performance Analysis

Table III shows the performance of the three mentioned algorithms on the three datasets. In case of the Adult dataset, filling the missing values with mean and mode resulted in model accuracies of 83.4% for KNN and 84.8% for logistic regression. After applying transformations KNN has an accuracy of 84.9% while logistic regression has an accuracy of 85.2%. F1-score for KNN is 0.6647 while for logistic regression it is 0.6599.

Table IV shows the results of different outliers handling techniques. It can be observed that the there is a significant improvement by removing outliers based on chi-square and dffits. There is not much of an increase in performance for the method of removing the wrongly classified outliers. But, in the datasets we used, the number of misclassified outliers are less. As a result, removing them did not make any significant difference. The non-parametric method is useful only when the variance in the variables is very high compared to the mean. There are few such variables in the Adult dataset and Mobile price classification dataset and hence we can see the improvement of performance in that case. For the EEG eye state dataset, the variance is very small compared to the actual values for all the variables, so there is no improvement in the performance of the model.

The results for the EEG eye state dataset stated are for KNN with a k value of 1. The leverage values are calculated based on the logistic regression model. In this case, the performance of logistic regression is low and hence we believe that the performance on this dataset did not improve with the handling techniques. However, the performance of this KNN model is slightly better than the algorithms used in [6] and [8].

V. Conclusion and Future work

Effective handling of outliers is important in terms of the model performance. This paper illustrates the performance of few outliers handling techniques and also the performance analysis of the popular machine learning algorithms KNN, Logistic regression and Random forest on three real world datasets. We observe that by using chi-square and dffits to remove outliers can produce significant improvement in the model performance. Non-parametric methods are effective when there are variables with high variance. It is also possible to implement more than one of these methods and we need to choose the methods which are suitable to the dataset. The future work can be based on the implementation of i) distance and clustering based methods when the performance of logistic regression is low, ii) handling missing values and iii) handling imbalanced data.

References

[1] E. Acuna and R. Caroline, "meta analysis study of outlier detection methods in classification", *Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez*, 2004

[2] H. Aguinis, R. Gottfredson and H. Joo, "Best-Practice Recommendations for Defining, Identifying, and Handling Outliers", *Organizational Research Methods*, vol. 16, pp. 270–301, 2008.

[3] A. Asuncion and D. J. Newman, *UCI machine learning repository*, 2007.

[4] V. Hodge and J. Austin, "A survey of outlier detection methodologies", *Artificial intelligence review*, vol. 22, pp. 85–126, 2004.

[5] M. I. Petrovskiy, "Outlier detection algorithms in data mining systems", *Programming and Computer Software*, vol. 29, pp. 228–237, 2003.

[6] O. Rösler and D. Suendermann, "A first step towards eye state prediction using eeg", in *Proceedings of the International Conference on Applied Informatics for Health and Life Sciences (AIHLS), Istanbul, Turkey*, 2013.

[7] M. R. Smith and T. Martinez, "Improving classification accuracy by identifying and removing instances that should be misclassified", *The 2011 International Joint Conference on Neural Networks, San Jose, CA*, pp. 2690– 2697, 2011.

[8] S. Vachiravel, "Eye State Prediction using EEG Signal and C4.5 Decision tree algorithm", *International Journal of Applied Engineering Research*, vol. 10, pp. 167–171, 2015.