A Methodical Approach for Dealing with Outliers in Classification

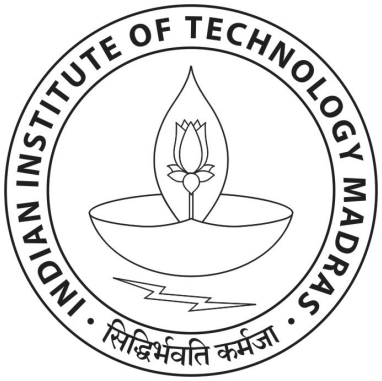
A PROJECT REPORT

submitted by

Vipul Spartacus Gandamalla
EE14B026

in partial fulfillment of the requirements
for the award of the degree of

BACHELOR OF TECHNOLOGY



DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS
MAY 2018

REPORT CERTIFICATE

This is to certify that the report titled **A Methodical Approach for Dealing with Outliers in Classification**, submitted by **Vipul Spartacus Gandamalla**, to the Indian Institute of Technology, Madras, for the award of the degree of **Bachelor of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Gaurav Raina
Project Guide
Associate Professor
Dept. of Electrical Engineering
IIT-Madras, 600 036

Place: Chennai
Date: 15th May 2018

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my guide Gaurav Raina for giving me an opportunity to work under him. Also I would like to thank you for constantly guiding me thoughtfully and efficiently throughout this project, giving me an opportunity to work at my own pace along my own lines, while providing me with very useful directions and insights whenever necessary.

I would also take this opportunity to thank all my friends who have been a great source of motivation and encouragement.

Finally I would also like to thank all of them who have helped me complete my project successfully.

Vipul Spartacus Gandamalla
EE14B026
Student
Dept. of Electrical Engineering
IIT-Madras, 600 036

Place: Chennai
Date: 15th May 2018

TABLE OF CONTENTS

Contents

1 INTRODUCTION **1**

2 Problem Setting **3**
- 2.1 Data Description . . . 3
  - 2.1.1 Adult dataset . . . 3
  - 2.1.2 Bank Credit Dataset . . . 3
  - 2.1.3 Diabetic Retinopathy Dataset . . . 4
- 2.2 Visualization . . . 4
  - 2.2.1 Adult dataset . . . 4
  - 2.2.2 Bank Credit dataset . . . 4
  - 2.2.3 Diabetic Retinopathy dataset . . . 4
- 2.3 Statistical Analysis . . . 4
  - 2.3.1 Adult dataset . . . 6
  - 2.3.2 Bank Credit dataset . . . 6
  - 2.3.3 Diabetic Retinopathhy dataset . . . 6

3 Outlier Management **6**
- 3.1 Defining Outliers . . . 6
  - 3.1.1 Error Outliers . . . 6
  - 3.1.2 Interesting Outliers . . . 6
  - 3.1.3 Influential Outliers . . . 6
- 3.2 Identifying Outliers . . . 6
- 3.3 Handling Outliers . . . 7

4 Algorithms/Models **7**
- 4.1 Logistic Regression . . . 7
- 4.2 KNN (K Nearest Neighbours) . . . 8
- 4.3 Performance Analysis . . . 8

5 Avenues for Future Work **8**

6 Conclusion **9**

7 References **9**

A Methodical Approach for Dealing with Outliers in Classification

Vipul Spartacus Gandamalla

15th May, 2018, Indian Institute of Technology Madras

Abstract

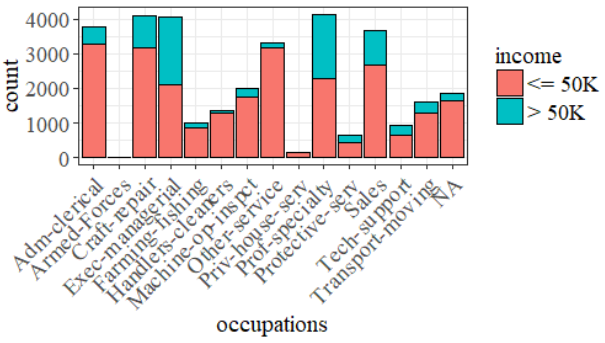
In this paper, we build machine learning models and discuss the effects the outliers have on these models. We give a formal definition of different types of outliers, discuss various identification techniques of those outliers and finally ways to handle them. The paper details in the exploration, preparing, modelling and evaluating the datasets and models. We applied transformations on various features based on their distributions. We consider two widely used machine learning algorithms for our predictions- (i) KNN(K Nearest Neighbours), and (ii) Logistic regression. To that end, we use three binary classification datasets (i) Adult dataset, (ii) Bank Credit dataset, and (iii) Diabetic Retinopathy dataset to understand the effect of the handling techniques on these datasets. We compare the performance of KNN and logistic regression on these datasets before and after handling the outliers using metrics like accuracy, F1 score, chi-square value. We observe that applying these techniques results in better accuracies for all above datasets.

1. INTRODUCTION

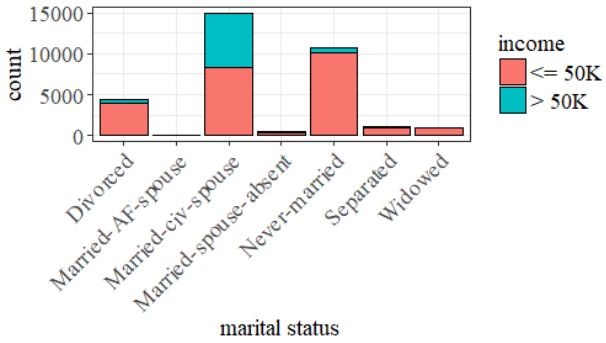
In this paper we investigate the effect the outliers have on models and their predictions. To that end, we use basic algorithms like KNN and logistic regression. These algorithms belong to a class which is simple to implement and understand and also works incredibly well in practice. They are surprisingly versatile for their simplicity. The objective is to build these KNN and logistic regression models on 3 datasets and evaluate their performance before and after handling the outliers. We observe that these comparisons are not universal and are dependent on the dataset at hand. Logistic regression is a binary classifier which models the input and gives the probability for the input to be true or false. KNN is a multinomial classifier which calculates the distance between the test sample and all the train samples and classifies the test sample based on the majority vote from k nearest train samples.

The 3 datasets we are working on are (i) Adult dataset (ii) Bank Credit dataset (iii) Diabetic Retinopathy dataset. These types of datasets can be used by governments to design budgets catering to a specific class of people or can be used by banks while providing loans. The diabetic retinopathy dataset can be used by hospitals to evaluate their test results. The three datasets were obtained from UCI machine learning repository, uploaded by [8, 9, 3].

The data has missing values in some of the variables, these are handled by observing the variation of the missing variables with respect to the target variable and the missing values are updated accordingly. KNN and logistic regression models are built after preprocessing the data. The rest of the paper is organized as follows: Section II give a brief description of our datasets, a few insights drawn from them, in Section III we give definitions of outliers, their identification and handling techniques. A brief introduction about KNN, logistic regression and performance analysis is given in Section IV. In Section V we give some avenues for future work, and we conclude our work in Section VI

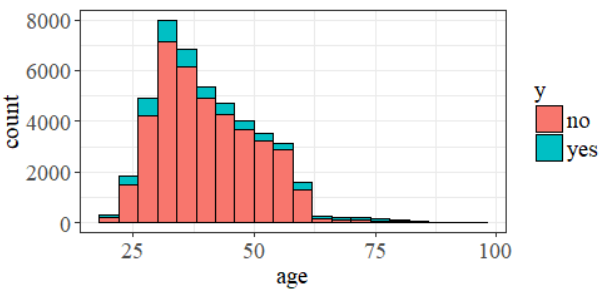


(a) Occupations with respect to income

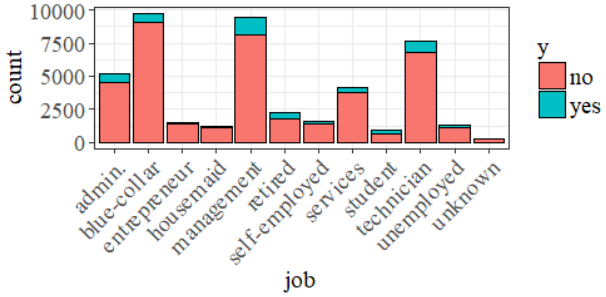


(b) Marital status with respect to income

Figure 1: Bar plots of variables with respect to income(target variable) for Adult dataset

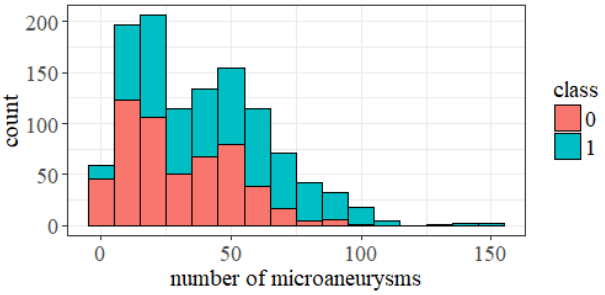


(a) Age with respect to y

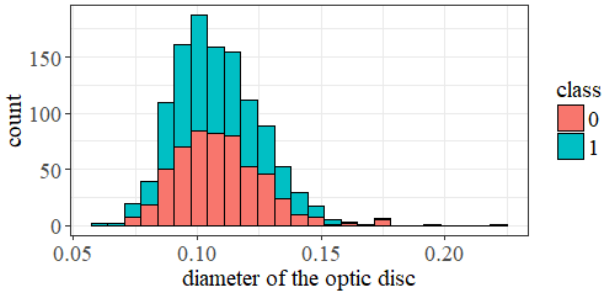


(b) Job with respect to y

Figure 2: Bar plots of variables with respect to y(target variable) for Bank Credit dataset

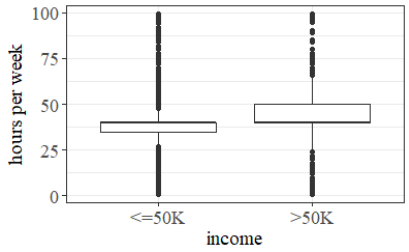


(a) Microaneurysms with respect to class

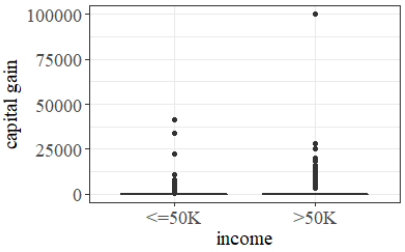


(b) Diameter of the optic disc with respect to class

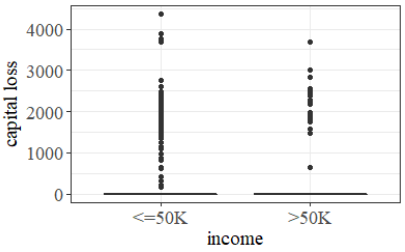
Figure 3: Bar plots of variables with respect to class(target variable) for Diabetic Retinopathy dataset



(a) Hours per week with respect to income

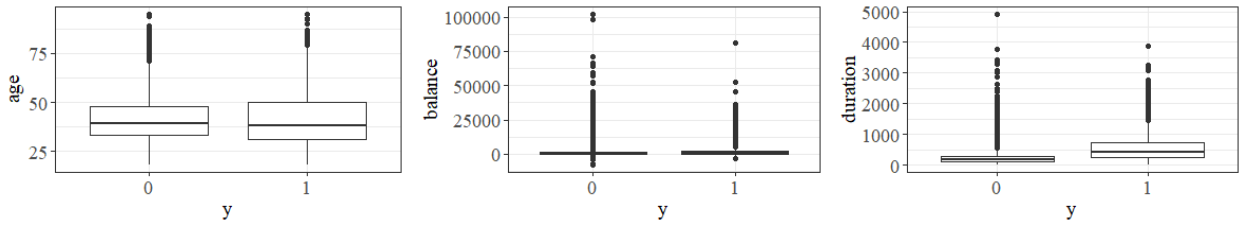


(b) Capital gain with respect to income



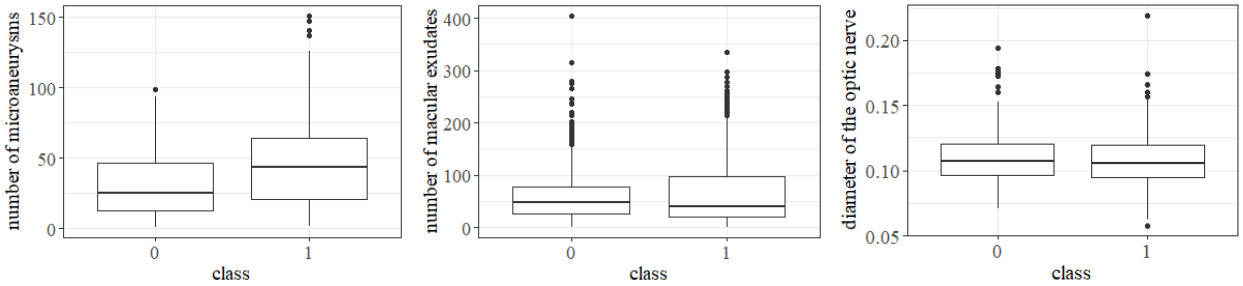
(c) Capital loss with respect to income

Figure 4: Box plots of variables in Adult dataset with respect to income(target variable)



(a) Age with respect to y (b) Balance with respect to y (c) Duration with respect to y

Figure 5: Box plots of variables in Bank Credit dataset with respect to y(target variable)



(a) Microaneurysms-a with respect to class (b) Macular exudates-a with respect to class (c) Macular exudates-e with respect to class(target variable)

Figure 6: Box plots of variables in Diabetic Retinopathy dataset with respect to class

2. Problem Setting

2.1. Data Description

2.1.1. Adult dataset

The Adult dataset has 32561 samples and each sample 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 numericaland 8 are categorical. 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 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.

2.1.2. Bank Credit Dataset

The dataset has 45211 samples or instances with each instance having 1 target variable and 16 feature variables. The feature variables include age, job, marital status, education, previous cases of default, etc. The objective is given these feature variables we have to suggest the bank if it is safe to provide loans to a new individual. This is a binary classification problem dependent on 16 variables out of which 7 are numeric and 9 are categorical. The dataset has missing values in variables job, education, contact, poutcome. The dataset is highly skewed with 11% of data belonging to one class and 89% to other.

Table 1: Correlations of variables in Adult dataset with income before and after the variable transformation

	work class	education num	marital status	occupations	race	native country
Before	0.003	0.460	0.564	0.047	0.121	0.031
After	0.211	0.477	0.611	0.477	0.137	0.10

2.1.3. Diabetic Retinopathy Dataset

Diabetic retinopathy is the leading cause of blindness in working age population. Currently detecting it is a time consuming manual process. The aim is to build an automated system that can detect this disease with high accuracy like the work done in [2, 10]. The dataset has 1150 instances with each instance having 1 target variable and 19 independent variables. The objective is to classify the test results as positive or negative with high certainty. All the variables in this dataset are numeric. The dataset has features like the diameter of the optic disc, microaneurysms found, etc. The dataset has no missing values. The dataset is almost balanced with 46% samples belonging to one class and 54% to other.

2.2. Visualization

2.2.1. 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. People in age group 40 to 55 tend to have income above 50K.

2.2.2. Bank Credit dataset

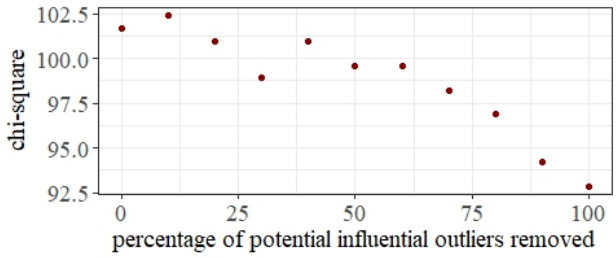
The dataset has 45211 samples of which 6851 people have studied primary, 23202 studied till secondary, 13301 have studied till tertiary education. The educational qualification of 1847 people is not known. The cellular or telephone contact details of 32191 people are known. The loan approval rate for people who provided their cellular contact is high. People who are administrators, have blue collar jobs or belong to management department or technician tend to have higher loan approval rate. All people with no previous record of default have their loan approved.

2.2.3. Diabetic Retinopathy dataset

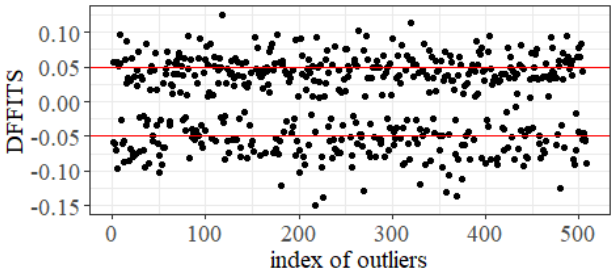
The dataset has feature variables like quality assessment, pre-screening, microaneurysms, macular exudates. The dataset has only 4 samples with bad quality. The pre-screening attribute which indicates retinal abnormality is not a good indicator of the disease as 50% of people with no retinal abnormality ended up having the disease. The nma.a-f features indicate the number of microaneurysms found in the eye. People with microaneurysms above 100 were found to have the disease irrespective of other attributes.

2.3. Statistical Analysis

Filling of missing values is a crucial part of any machine learning problem. The filling has to be done in such a way that it preserves the initial distribution of the data. Generally, for filling the missing values we need to observe the mean, outliers, median in case of continuous variables. 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.

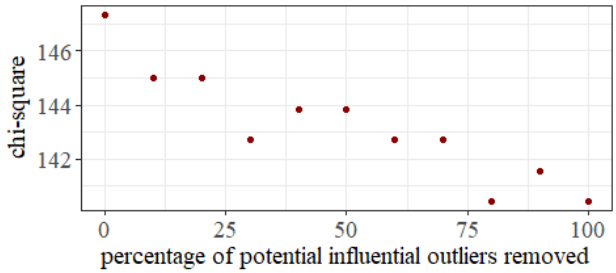


(a) Variation of chi-square with removing potential influential outliers

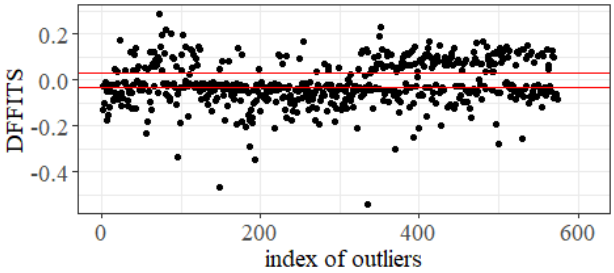


(b) Threshold for DFFITS values of outliers

Figure 7: chi-square and DFFITS values for Adult dataset

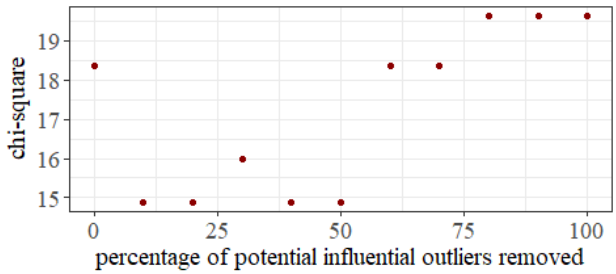


(a) Variation of chi-square with removing potential influential outliers

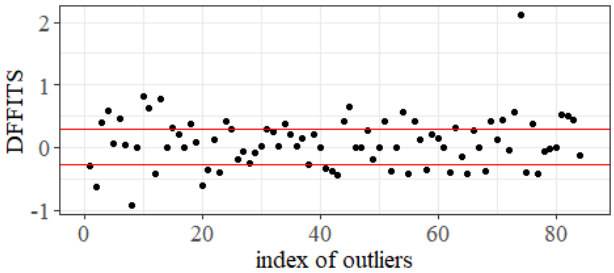


(b) Threshold for DFFITS values of outliers

Figure 8: chi-square and DFFITS values for Bank Credit dataset



(a) Variation of chi-square with removing potential influential outliers



(b) Threshold for DFFITS values of outliers

Figure 9: chi-square and DFFITS values for Diabetic Retinopathy dataset

2.3.1. Adult dataset

It is observed that the data has missing values in workclass, occupations and native country. 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. The correlations have increased significantly after above transformations as can be seen in Table I for some variables thus increasing the accuracy and F1 score.

2.3.2. Bank Credit dataset

The dataset has missing values or unknowns in variables like job, contact, poutcome. In this problem, we observe the variation of these variables with the target variable and merge accordingly. The variable job has 288 missing values these are merged with housemaid as this class of job follows almost same distribution as missing values. Similarly, in the variable contact, the missing values are merged with cellular, in the variable poutcome, the missing values are merged with other.

2.3.3. Diabetic Retinopathhy dataset

This dataset doesn't have any missing values but is highly skewed for variables like quality assessment, pre-screening. The variable diameter of optic nerve closely follows a normal distribution, therefore, we scale this variable. Similarly, the variable dd also follows a normal distribution and it is scaled. The transformed values are used for fitting the model.

3. Outlier Management

We follow [1] in defining, identifying and handling outliers. These methods entail a clear transparent approach towards dealing with outliers.

3.1. Defining Outliers

3.1.1. Error Outliers

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

3.1.2. 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.

3.1.3. 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.

3.2. Identifying Outliers

Identifying outliers is done using various single construct and multiple construct techniques. In single construct techniques we use box plot as mentioned in [4, 7] and percentage analysis to identify potential error outliers. In multiple construct techniques, we use leverage values. Leverage values are the diagonal values of the hat matrix or projection matrix. They indicate how far away independent predictors of an observation in the feature set are away from rest of the observations.

In single predictor datasets extreme value of the predictor indicates outliers as well as high leverage points, but in multiple predictor datasets, high leverage points are those that have either extreme independent predictor variables or an unusual combination of multiple predictor variables. The cutoff value for the leverage values is $2(k+1)/n$, where k is the number of predictors and n is the sample size. The points beyond this cut off are potential error outliers.

Potential error outliers should be cross verified against original records to declare them as error outliers. In our case this is not possible, hence we do not consider these potential error outliers as error 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.

To identify model fit outliers, we observe the effect of each potential influential outliers on the model fit [7]. 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. As the data size is huge, the classification doesn't change just by removing one point. The potential influential outliers are inserted in sets in the remaining data and the changes in chi-square are plotted. The set of outliers which when included provides the least chi-square value is taken and the final model is built using this set. If the chi-square value doesn't change during this operation the potential influential model fit outliers become potential influential prediction outliers.

For potential influential prediction outliers, we observe the effect of each potential influential prediction outliers on parameter estimates. This is done by calculating Cooks distance [6], DFFITS [5] which measures the difference in the estimated outputs of the models with and without the data point. They show how influential the point is in the analysis. Threshold value is indicated by the red line in Figures 7b, 8b and 9b for prediction outliers is $2\sqrt{(k+1)/n}$. The points that lie beyond this threshold are prediction outliers.

3.3. 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 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 class of the outliers is predicted and the wrongly classified outlying points are removed from further analysis.

4. Algorithms/Models

4.1. 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.

Table 2: Accuracies obtained for three datasets after applying handling techniques

	Adult dataset	Bank Credit dataset	Diabetic Retinopathy dataset
Keep them	84.9%	90.51%	77.39%
Remove Them	85.8%	90.38%	75.11%
Remove wrongly classifed	85.1%	90.12%	78.53%

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

4.2. KNN (K Nearest Neighbours)

KNN is a non-parametric lazy learning algorithm. Non-parametric means that it does not make any assumptions on the underlying data distribution. 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 neighbours. A test sample is classified by the majority vote of its neighbours.

4.3. Performance Analysis

With our definitions of outliers and their identification techniques, we obtained leverage values of all the samples and using these leverage values we arrived at Figures 7,8,9. Using the Figures 7a, 8a and 9a we removed the percentage of outliers that resulted in least chi-square value. From Figure 7a, the least chi-square value is obtained when we remove all the outliers, this provides the best accuracy for adult dataset as can be seen in Table II. For adult dataset removing all outliers gave better results, this could imply that all the outliers are error outliers, as we do not have access to original records we cannot confirm this. From Table II it can be observed that bank credit dataset has highest accuracy when we keep all the outliers as opposed to the insights from the Figure 8a. From Figure 9a, when 10% of outliers are removed we get least chi-square value, as this is almost equivalent to not removing any outliers we get maximum accuracy when we keep all the outliers as seen in Table II. We observe that if the plots in Figures 7a, 8a and 9a had constant values we would have to use DFFITS values to identify influential outliers.

5. Avenues for Future Work

The thresholding techniques used in Cook's distance and DFFITS are not universal, they are dataset and objective dependent. Way to identify these thresholds should be looked into. While calculating DFFITS we remove a potential outlier, fit a model on rest of the data and predict this outlier. The aim is to find if this outlier is causing a huge shift in the model. This assumes that only one outlier is present in that vicinity. If a bunch of outliers are present near this outlier, removing this one doesn't shift the model, so it might be wrongly classified as not an outlier. Therefore before applying DFFITS, we need to cluster the outliers, and whenever removing an outlier to fit a model we need to remove all the outliers from its cluster.

6. Conclusion

Outliers are observations that deviate markedly from rest of the data, influencing model fit. In this paper we give a clear definition of different types of outliers, systematic identification techniques of these outliers and some of the handling techniques. The techniques were implemented on three datasets, (i) Adult dataset, (ii) Bank Credit dataset, and (iii) Diabetic Retinopathy dataset. We observe that the results of these applications are highly dependent on the dataset. Further analysis can be done on outliers obtained by using the above techniques to gain valuable insights about the data.

7. References

[1] H. Aguinis, R. Gottfredson and H. Joo, "Best-Practice Recommendations for Defining, Identifying, and Handling Outliers", *Organizational Research Methods*, vol. 16, pp. 270–301, 2013.

[2] M.U. Akram and S.Khalid and S.A. Khan, "Identification and classification of microaneurysms for early detection of diabetic retinopathy", *Pattern Recognition*, vol. 46, pp. 107–116, 2013.

[3] B. Antal and A. Hajdu, "An ensemble-based system for automatic screening of diabetic retinopathy", *Knowledge-Based Systems*, vol. 60, pp. 20–27, 2014.

[4] V. Barnett and T. Lewis, *Outliers in Statistical Data*,John Wiley & Sons, 1994.

[5] D.A. Belsley, K. Kuh and R.E. Welsch "Regression diagnostics: Identifying influential data and sources of collinearity", *Journal of Applied Econometrics*, vol. 4, pp. 97–99, 1980.

[6] R.D. Cook, "Detection of Influential Observation in Linear Regression", *Technometrics*, vol. 42, pp. 65–68, 2000.

[7] J. Han and M. Kamber, *Data Mining : Concepts and Technique*,Morgan Kaufmann, 2012In this paper we

[8] R. Kohavi and B. Becker, "UCI machine learning repository", 2014.

[9] S. Moro , P. Cortez and P. Rita, "A Data-Driven Approach to Predict the Success of Bank Telemarketing", *Decision Support Systems*, vol. 62, pp. 22–31, 2014.

[10] D. Usher , M. Dumskyj, M. Himaga, T.H. Williamson, S. Nussey and J. Boyce, "Automated detection of diabetic retinopathy in digital retinal images: a tool for diabetic retinopathy screening", *Diabetic Medicine*, vol. 21, pp. 84–90, 2013.