7CS516 Processing Big Data

Find smokers by vital signs

Ntoulmperis Michail[1]

[1]107 Patision anenue & Pellinis Str, 11251, Athens, GR
UoD: 100615926, mich.ntoulmperis@mc-class.gr
MSc Big Data Analytics

Abstract. For this study we use four machine learning algorithms on a dataset containing vital measurements of smokers and non-smokers. While in essence we faced a binary classification problem, meaning each observation can be classified into two classes, we used a variety of machine learning algorithms to see how much accurate they can be even when they are not used for classification. The program language used was R in the environment of RStudio. During the coding process we tried to optimize some the algorithms used to get better results. After the application of these algorithms, we evaluated them and reached a conclusion on which seems the best and why.

Keywords: Smoking, Evaluation, Machine Learning, R, Classification, Linear Regression, Logistic Regression, Support Vector Classifier, Principal Component Analysis.

Contents

1 Introduction

This study is revolved around a data set obtained from Kaggle [1] which includes different physiological measurement taken from smokers and non-smokers. These variables can be given meaning though various of interpretations, meaning the gender or the height of each person maybe can greatly influence their decision they to smoke[2]. On the other hand, physiological measurements such as blood pressure for example can give us an idea on how much it is affected through smoking. What we mean by this, is that if for example blood pressure is statistically significant in a model that predicts if a person smoke or not, then we can assume that there is a great difference in values in smoker and non-smokers.

In order to meet the objectives above we are going to use four algorithms from the textbook [3]. To begin with, we will try and build a model of Linear Regression using the "smoking" variable as a response and will try and find out how it performs even if this isn't a data set suitable for regression. We then are going to apply a Logistic Regression model. which is a machine learning algorithm suitable for classification as it computes the probability of an observation to be in selected class. Furthermore, support vector classifier is going to be used, as they it is a famous classification algorithm and one that we can get great visual cues and give us great insight on what way the classification actually looks like. Last but not least, we are going to use and algorithm in the category of Unsupervised Machine learning. Principal Component Analysis is going to produce a low-dimensional representation of the dataset of our study. We will use it to find multiple linear combinations of the variables that have the maximum variance and are uncorrelated with each. Lastly, we will use PCA as a visualization tool.

In the next part of our study, we are going to explain how the preprocessing of the data set was done in order to make it suitable for analysis.

2 Data Presentation and Pre-Processing

As mentioned above out data set contains measurement from smokers and non-smokers[1]. Our data set has 55692 observations and 27 variables. During the coding process we made sure no missing values were included in our data because that will lead to a great amount of uncertainty, meaning it will alter our computation of estimates

and metrics of each model[4]. Another part of preprocessing is going to be deleting any unwanted data, with this we save us processing power, produce faster and more reliable results as we will make sure that all data used to train the machine learning algorithms is useful to the model, maximizing its efficiency and accuracy. We removed from the data set the first column which included the ID number of each observation as it will serve no purpose to our analysis. Another variable that was deleted from the data set was the "oral" variable which included the oral status from each person, the reason of the deletion was this column had in its entirety the "True" logical value and it wasn't going to contribute to our analysis. Lastly the values of "gender" and "tartar" variables were given the "1" or "0" value depending on their original value. This was done in order to make its easier apply machine learning without worrying if the variables are logical or character. The final data dataset had the same number of observations as the one we started and 25 variables. A description of the variables is given in **Appendix A**.

3 Algorithms

3.1 Linear Regression

For this algorithm, the formula is one of the parameters you can adjust to affect your prediction model. Linear Regression can be used in conjunction of the subset function, which finds the optimized model through doing multiple calculations of the coefficients while having as an input all the variables. In simpler terms, we first build a model with all the variables and then selected which are statistically significant and then build another with only those. This way we save processing power and can greatly improve the accuracy of the model. After the model is done, we have built a linear equation that predicts the response. Coefficients and specifically p-value is the one we should care the most since if p-value is more than 0.05 it means the coefficient is statistically not significant and irrelevant to the model[3]. R Squared is a metric that explains the proportion of the variance that can be explained. The range of R Squared is from 0 to 1. Other important metrics are RSS(Residual Sum of Squares) , Cp and Bic (Bayesian Information Criterion) which are used to assess the fit of a regression model that has been estimated using ordinary least squares.

3.2 Logistic Regression

Logistic Regression is mostly used in binary classification problems. A big difference between this algorithm and Linear Regression is that with this algorithm rather than trying to predict the value of the variable "smoking" we are interested in predicting the probability of it being in the class "1" or "0". In order to train a logistic regression model in R we will use the built in generalized linear models, $glm()$ function. The syntax is similar to that of $lm()$, except we need to pass the argument family=binomial in order to tell R to run a logistic regression model. As far as metrics go, we will mostly discuss the same ones we did in Linear Regression. Since Logistic Regression is suitable for binary classification, we also made sure to split the data set into

test and train datasets. We also made a prediction and tried to calculate the accuracy of the model using the $table()$ command.

3.3 Support Vector Classifier

The support vector classifier as an algorithm is great for a binary classification problem and it can provide great visualization of how the classification was done. We get our results from enlarging the feature space in a specific way using the $kernel = "linear"$. Then, using the $svm()$ command we generate a "svm" model. In this stage we can select how strict the algorithm is going to be and how difficult and detailed the computation of the separation of classes is going to be by selecting a value for the cost. This selection is not of great important since with the tune command we can use cross validation to find the optimum value for the cost parameter. After the optimal model has been found we can plot out result to see how the classification was done by the algorithm and through the $summary()$ command we can see how many support vectors were created.

3.4 Principal Component Analysis (PCA)

The last algorithm used in our report is Principal Component Analysis (PCA), is an algorithm that belongs to unsupervised machine learning and that has a built-in function named prcomp. In essence it simplifies the complexity in high-dimensional data while retaining trends and patterns. It does this by transforming the data into fewer dimensions, which act as summaries of features[3]. PCA reduces data by geometrically projecting them onto lower dimensions called principal components (PCs), with the goal of finding the best summary of the data using a limited number of PCs. There are parameters that potentially can affect the result of the analysis such as the "scale" which indicate whether the variables should scale before analysis to gain unit.

4 Application – Evaluation

To start with, as mentioned we firstly built a model of Linear Regression. As seen in Appendix B in the residuals plot, we can determine that the model has a very bad performance as expected. We tried to improve its performance with the $subsets()$ command in order to find a subset of variables with the most statistically significant coefficients. After that by using the $summarry()$ command we found what the Intercept and those variables were by reviewing plot of Bic, CP, R Squared and Adjusted R Squared, where 16 of 25 variables where selected. To sum up the R Squared of the model was 0.3033785 which shows us how bad performance the model has and further proves Linear Regression should not be used for classification.

On the other hand, Logistic Regression which is advised for binary classification proves to be a better fit. First, we spitted the data into test data set and a train data set. We used the "best" variables found from Linear Regression and train a model, then we made some plots to showcase its performance as see in **Appendix C**. As we can see in

comparison of the residual plots, Regression was had way fewer residuals. After that we made a prediction using the $prediction()$ command and made a confusion matrix from which we computed the accuracy of the model. The respective accuracy and error is 0.7459177 and 0.2540823 which shows great performance.

In the application of the Support Vector Classifier(SVC), we found the correlation of the variables and selected those with correlation great that 0.2 in order to minimize the computational cost. To further reduce the computational cost a random sampling of the data set was made, reducing its size from approximately 55.000 rows to 1241 rows along the 25 variables for the train dataset, which was named "ssndf_train" Then using cross validation we saw the optimum cost for our model was $c = 10$. Finally, through the $table()$ and $predict()$ commands we produce a confusion matrix the informs us on the accuracy of the model. The accuracy of the model was not good, and maybe the continuous sampling was at fault.

For PCA we started by computing the respected Principal Components. After that we made a plot of the first two Principal Components as seen in **Appendix E**. We can see that most variables are separated into two clusters which can give us insight for future analysis. Then we tried to make plots of PVE(Proportion of Variance Explained) as seen in **Appendix F**. We can observe that as more Principal Components are Calculated the less the proportion of their variance is explained.

5 Conclusion

In conclusion, after the proper preprocessing was done and the filtration of data, meaning deleting columns and transforming the types of other into the most suitable type for analysis we proceeded to the machine algorithms. In the future we suggest removing more outliers and leverage points. Throughout the process of building different models, we found out the Linear Regression is definitely not a suitable algorithm for binary classification even when finding the best subset of variables. With Logistic Regression we managed to build a model of great accuracy and minimal error, thanks to splitting the dataset, ensuring an unbiased model got created, while also utilizing the subsets acquired from the previous algorithms.

By using the Support Vector Classifier algorithm, due to its huge computations and the size of our data we didn't manage to get a plot that shows how the classification was done visually. Even when we selected only the highly correlated values and greatly reduced the size of the dataset, we couldn't get viable results. If the dataset had smaller dimensions or if unsupervised machine learning algorithms had been applied during preprocessing, we may had had different results.

PCA helped us reduced the dimension of our dataset by computing several distinct Principal Components. These PC can help us reduce the computational cost although this method must be used with caution since tit can sometimes decrease the accuracy of the rest of the models. However, we managed to visualize how the Proportion of Variance Explained changes though the computation of the Principal Components and we also managed to plot the first two PC that shows that the data set seems to be

organized into two large clusters as we saw in **Appendix D**. To sum up, Logistic Regression was by far the most efficient algorithm and for future analysis we suggest to further use PCA before the modeling so we can achieve dimensionality reduction. That way the computational cost of all the models applied in our study would be greatly decreased and more accurate result and plots will be made.

References

[1] Kukuroo3, "Find smokers by vital signs," *https://www.kaggle.com/*. https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking (accessed May 23, 2022).

[2] N. Experiments, "Editorial Identification of Treatment Effects," vol. 1131, no. 2007, pp. 1127–1131, 2008, doi: 10.1002/hec.

[3] T. Hastie, R. Tibshirani, G. James, and D. Witten, "An introduction to statistical learning (2nd ed.)," *Springer texts*, vol. 102, p. 618, 2021.

[4] X. Wang and Y. He, "Learning from Uncertainty for Big Data," *Ieee Syst. Man Cybern. Mag.*, no. August, 2016, [Online]. Available: http://www.hebmlc.org/UploadFiles/20161121203535376.pdf.

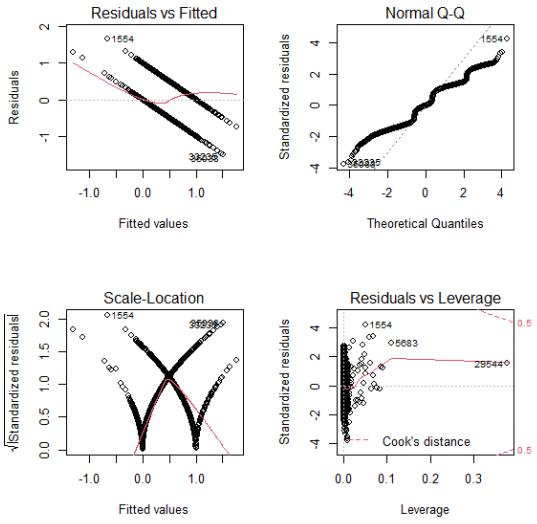
Appendices

Appendix A

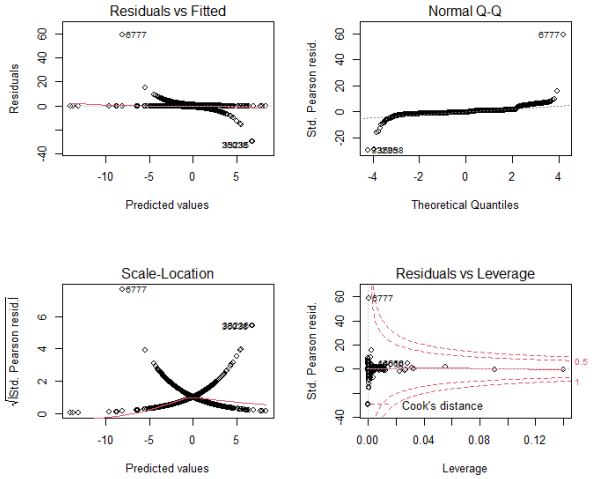
Description of the Variables.

- **gender :** If "1" the observation was male and "0" it was female
- **age :** The age of each observation
- **height.cm. :** The height in cm of each observation
- **weight.kg. :** The weight of each observation in kg
- **waist.cm. :** The waist size of each observation in cm
- **eyesight.left. :** A score of that describes each observations eyes sight in its left eye
- **eyesight.right. :** A score of that describes each observations eyes sight in its right eye
- **hearing.left. :** A score of that describes each observations' hearing in its right ear
- **hearing.right. :** A score of that describes each observation hearing in its left eye
- **systolic :** Blood pressure measurement of each observation.
- **relaxation :** Blood pressure measurement of each observation.
- **fasting :** Blood sugar measurement of each observation.
- **Cholesterol :** Total Cholesterol measurement of each observation.
- **triglyceride :** Triglyceride measurement of each observation.
- **HDL :** Cholesterol type measurement of each observation.
- **LDL :** Cholesterol type measurement of each observation.
- **hemoglobin :** Hemoglobin measurement of each observation.
- **Urine protein :** Urine protein measurement of each observation.
- **serum creatinine :** Serum creatinine measurement of each observation.
- **AST :** Glutamic oxaloacetic transaminase type measurement of each observation.
- **ALT :** Glutamic oxaloacetic transaminase type measurement of each observation.
- **Gtp :** γ-GTP measurement of each observation.
- **dental caries :** Number of dental carries of each observation.
- **tartar :** If "1" the observation had tartar and "0" it had not.
- **smoking :** If "1" the observation was a smoker and "0" it was a non-smoker.

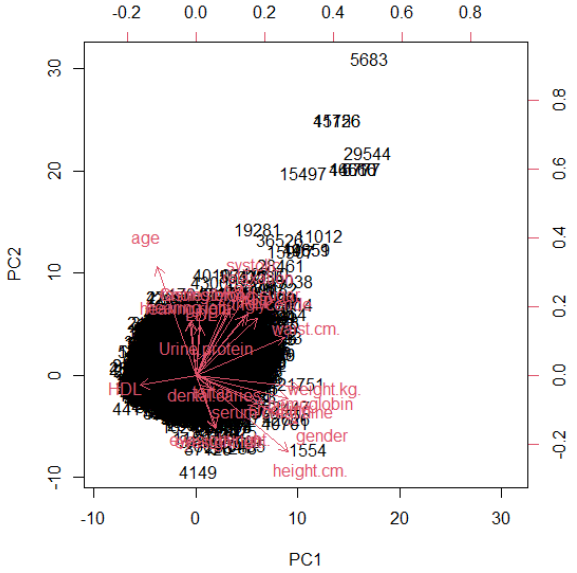
Appendix B



Appendix C



Appendix D



Appendix F