CAPSTONE PROJECT – STATISTICAL DATA MINING FOR BIG DATA

**Using Machine Learning Methods to Detect Autism Spectrum Disorder in Adults**

**William Belcher**

Autism Spectrum Disorder is a neurodevelopment condition in which the people with the condition exhibit a variety of distinct behavioural patterns. Diagnosis typically happens at a young age as parents and teachers notice the autism symptoms. However, identifying autism in adults is particularly difficult as its symptoms overlap with a variety of other mental health conditions. With machine learning techniques growing ever popular in the clinical setting, a rapid screening process that could be used in assisting the referral of patients to medical professionals was essential. In this paper, the possibility of using Logistic Regression, K-Nearest Neighbours and Naïve Bayes classifiers is explored. The proposed methods were evaluated on a publicly available AQ-10-Adult based screening data set that contained 704 observations of 21 attributes. After the relevant pre-processing of the data and application of the above-mentioned techniques, the achieved results indicated …

**Introduction**

Autism Spectrum Disorder (ASD) is a neurological condition associated with many atypical mannerisms and behavioural patterns, most notably those surrounding interpersonal interactions. Autism Spectrum Disorder is a condition related to the development of the human brain. It is worth noting that both environmental and genetics may be contributing factors in the development of ASD. However, scientists have been unable to uncover the root cause and as such ASD is usually detected through observations and diagnosed by a specialist. Unfortunately, the process for receiving an ASD diagnosis are lengthy with multiple appointments with specialists not being cost effective. Early detection of the condition can assist in the improvement of the subject’s overall health by enabling them to implement techniques and medication that reduce the impact of their condition on their daily lives sooner. With the rapid increase in modern computing power and number of machine learning models assisting in the diagnosis of medical conditions, the early detection of ASD based on variety of physiological attributes now seems viable. The detection of autism spectrum disorder in a patient proves difficult as, as the name implies, the disorder is a spectrum resulting in significant intragroup variance in those being classified as having the condition. A time-efficient and easily accessible screening process is necessary in assisting medical professionals in informing individuals whether they should pursue a formal, clinical diagnosis.

**Data**

The Autism Screening Data for Adults data set, collected from the UCI Machine Learning Repository, contains several predictors and one target variable [https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult]. The attribute types for these predictors are either categorical, continuous or binary with the response variable being a categorical “yes” or “no”. The dataset’s 20 attributes used for prediction are listed below:

|  |  |
| --- | --- |
| **Attribute ID** | **Description** |
| 1-10 | Answer to the corresponding AQ-10-Adult [] question |
| 11 | Patient age |
| 12 | Gender |
| 13 | Ethnicity |
| 14 | Did the patient have jaundice at birth? |
| 15 | Family history of ASD |
| 16 | Patient country of residence |
| 17 | Has the patient used the screening app before? |
| 18 | Screening score |
| 19 | Age group |
| 20 | Person who’s using the screening app’s relation to patient |

Table 1: Attribute list

In total, there were 704 observations. It is worth noting that the screening score is the sum of the answers to the AQ-10 questions, where a 1 represents a ‘slightly agree’ or ‘strongly agree’ and a 0 represents a ‘slightly disagree’ or ‘strongly disagree’.

*Data Pre-processing*

Data pre-processing is a technique in which the raw data is transformed into a meaningful and understandable format. The ‘Age group’ column was removed from the data as it only contained one value of ’18 and over’. The ‘Screening score’ column was also removed as it was just the sum of the answers to the AQ-10 questions. Missing values in the data are denoted with a ‘?’ with rows containing missing values being removed from the data. Missing value imputation was not conducted in order to not introduce unnecessaryvariance in the predictors. A single outlier was detected and remove. The size of the cleaned data used for model generation contained 19 columns, 18 predictors and 1 response, with 608 total observations.

**Methods**

*Principle Component Analysis (PCA)*

Principal component analysis is an unsupervised learning technique used for dimension reduction as well as exploratory data analysis by projecting the variables onto a new, orthogonal basis that can be used to illustrate the proportion of variance explained by each principal component.

*Training and testing split*

The complete data set was split into training and testing subsets using an 80/20 split. K-fold validation with the training subset and with k = 5 was used for each model. Figure 1 shows the final training, validation and testing sets that were used.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Complete Dataset | | | | | | |
|  | | | | | | |
|  | | | | | | |
| Training Set (80%) | | |  | Testing Set (20%) | | |
|  | | | | | | |
|  | | | | | | |
| 1 | 2 | 3 | | | 4 | 5 |

Figure 1: Data split

*Logistic Regression (LR)*

Logistic regression is based off the standard linear regression methods but applies the logit transformation to the resulting in a response that will lie between the value of 0 and 1 inclusive.

*K-Nearest Neighbour (KNN)*

KNN is a supervised learning approach that can be used for both regression and classification problems. The KNN classifier used here solves classification problems by estimating the conditional probability as the average of the training responses in the neighbourhood. The distance measure used is the Euclidean distance.

*Naïve Bayes*

The naïve Bayes classifier is a probabilistic machine learning model that is based on Bayes theorem, finding the probability of happening given that has already occurred. The naivety of the model comes from the assumption that features are independent in each class.

**Results and Discussion**

*Performance Metrics*

The result of a model is measured in terms of its specificity, accuracy and sensitivity. The values are obtained by using the generated models to predict the outcome on the test data set and making note of the resulting confusion matrix.

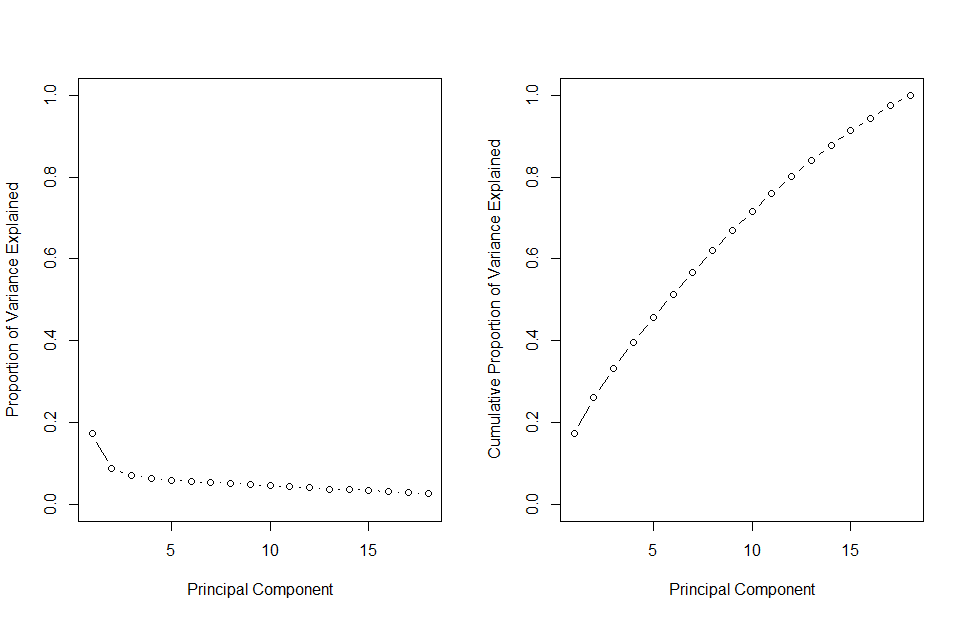
|  |  |  |
| --- | --- | --- |
|  | Predicted Outcome | |
| Actual Outcome | True Positive | False Positive |
| False Negative | True Negative |

Table 2: Confusion Matrix

|  |
| --- |
| Performance Metrics |
|  |
|  |
|  |

Table 3: Performance Metrics formulae

Principle component analysis illustrated that each principle component is necessary in explaining the variation within the data with PC1 explaining 17.28% of the variance all the way through to PC19 which explained 2.5%. The resulting figures showed no distinct ‘elbow’ outside of PC1.



**Conclusion**

When looking at the decision importance tree, as well as figure 2, figure 3, figure 4 and figure 5, it can be seen that posting a video will drastically improve the engagement efficiency of the post. Posting in spring will, on average, bring a 20% increase to the engagements received. Altering the weekday and hour a status is posted provide minimal benefits when compared to the status type and season. If one were so inclined, posting on either a Monday or Tuesday between the hours of 6 to 10am will provide an approximate 45% increase in engagements over a post during any other time period.

In conclusion, the best combination of post type and time posted that will maximise engagements is a Video posted between 6 to 10AM on either a Monday or Tuesday.

**References**

* [4] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA <http://www.rstudio.com/>

Mode: Desktop. Version 1.25042