4/7/2019

Shubham handa

8638369

ASSIGNMENT 4

MATHEMATICS FOR DATA ANALYSIS

Q1)

1. The first Step of any analysis project is to find the data and then transform the data to make analysis. Some of the data in the data autism character so they were transformed into numeric values. Variables like ethnicity, Rel, Autism, gender, Jaundice were transformed using assigning them numeric categorical values. This also helped in treating na values as it was given 0 value. In the case of age, we cannot replace the null value with any value value so it was replaced with the middle value i.e median value. Also, yes and no were divided into two categories i.e 0 for no and yes was 1 in each required variable.

Q2)

1) The next step is to do descriptive analysis; all the univariate graphs were plotted such as

Histogram, boxplots and scatterplot. Barplot was also plotted but it showed that it was

Not a good graph choice

1. The correlations were calculated using multiple correlation variable. The graphs were plotted according to that (using multiple correlation method). The method was chosen spearman. Also, each correlation was analyzed against autism and the variables which had the highest effect on autism were analyzed. These variables will be the guideline for all model building.
2. Treatment of outliers is a very sensitive thing, most of the data was categorical so it was not treated. For eg, I assigned the value in ethnicity pasifika as 9. It will be always outlier as compared to other values (as not occurring often). The only value which could be treated was Age. The boxplots were plotted to evaluate and identify which values could be outliers. There was one value in age which was highly off (383) this was replaced by using a technique whereas the values above than 98% quantile were replaced with 55 i.e 98% quantile value.
3. The thing about treating outliers varies data to data i.e you need to closely evaluate data sometimes the data might be correct) in that case you cannot just remove or treat the data. Also, for some missing values in 2 cases you need to either remove them or replace them with null values. There was one value in age which was highly off (383) this was replaced by using a technique whereas the values above than 98% quantile were replaced with 55 i.e. 98% quantile value.
4. A general idea was evaluated that which variables will have the highest impact on autism by looking at the correlations. These were then plotted graphically to prove and statistically prove it. Then I created a forward model and backward model. These were done by evaluating independence of each variable with autism by running chisq test and running and also plotting contingency plots.

Q3)

1. I Created 6 models in my process to identify and predict autism. The models were baseline which consisted of all the variables. After this the next model was to create stepwise model. I was not satisfied and I thought I could improve my model, so I started manually removing variables till I got my desired model. This was done by removing which don’t have high impact on autism such as x and x.1 and rel and ethnicity. It was again proved that variables such as res and a09 a06 a05 have the highest impact. But while evaluating my model the factor res was influencing autism in a very unusual manner i.e it was highly affecting the data (only true negatives and positives if we select res). So, in order to make my model more predictive and truer I decided to drop that too.

2) The statistics were justified and each model was plotted. The AIC value for each model was evaluated i.e it should be the lower. Also, deviance was evaluated to choose the minimum and the co-efficient were evaluated. It was observed again the model which was removed the values x and x.1 res ethnicity etc. The different models were used to get an idea between automated and manual methods. This helps us identify which variables can be removed.

3) The model that I chose is highly accurate, and the most significant factors are a09 a06 a05 and res. The variation it explains by choosing multiple variables i.e a lot of factors can result to autism. It is not overfitted to only one specific factor

Q4)

1. The next step after modelling was to do evaluation and classification in terms of our model. Two of the best models were chosen I.e general baseline and the model which was manually created were chosen. Predictions of model was evaluated and Confusion matrix was plotted. Again, predictions were made on the model to evaluate how good it is. Also, the Roc curve was plotted for each and auc was calculated highest being the best
2. On basis of the Confusion matrix specificity precision accuracy was calculated. Out of the three the one that I would recommend is accuracy because it has the highest impact in my opinion while evaluating.
3. The model I chose was manually created because it had the highest auc and less aic. Also, it had the least deviance and the coefficients were also better. The transformations made were dividing each variable into categories i.e categorical variable and the values. The main aim was to build a model to predict autism and what variables affect autism.it was observed that factors A05, A06, Res, A09 had the highest impact, but sometimes some variables might highly affect the model so they were removed. the aic value and auc values were minimum and close to 1 respectively. Hence stating that the model built was good and can be used for future predictions