

**ANL307e**

**Predictive Modelling**

**Group-based Assignment**

**TG02-Group 16**

**Team Leader:**

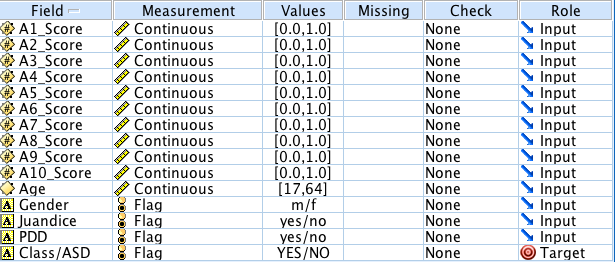
**Peter Reuben Liem (Q1581152)**

**Team Members:**

**Ng Wei Hao (Q1581197)**

**Tham Yun Zhi (B1581038)**

**a)**



*Figure 1: Data measurements and roles of dataset*

Data measurements are distinguished according to the types of values of the data. Data types can be classified as Continuous, Categorical, Flag, Nominal, Ordinal and Typeless. Numeric values with a range are set as “Continuous”. String data are set as “Categorical”. Data with two distinct values are set as “Flag”. Data with multiple distinct values are set as “Nominal”. Data with multiple distinct values in an inherent order are set as “Ordinal”. Data not used in the modelling process are set as “Typeless”.

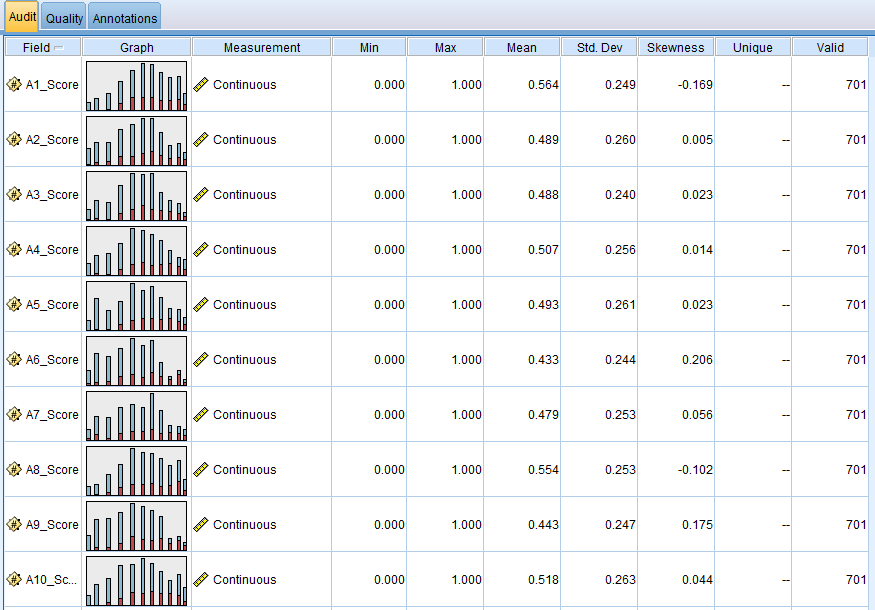
The measurements for A1\_Score, A2\_Score, A3\_Score, A4\_Score, A5\_Score, A6\_Score, A7\_Score, A8\_Score, A9\_Score, A10\_Score and Age are set as Continuous given that the numeric values are in a range. All 10 scores have a range from 0.0 to 1.0 and Age has a range from 17 to 64. On the other hand, the measurements for Gender, Juandice, PDD and Class/ASD are set as Flag given that the data are two distinct values. Gender has two distinct values of “m” and “f” and Juandice, PDD and Class/ASD have two distinct values of “Yes” and “No”. None of the fields are measured as Categorical, Nominal, Ordinal or Typeless.

Data roles are distinguished according to how the variable will be used in the modelling. Data roles can be classified as Input, Target, Both, None, Partition, Split, Frequency or RecordID. Variables used as an input are set as “Input”. Variables used as output or target are set as “Target”. Variables used as both input and output are set as “Both”. Variables that have no role assignment are set as “None”. Variables used to partition data into separate samples for training, testing and validation are set as “Partition”. Variables that specifies a model to be built for each value of the field are set as “Split”. Variables that are used as a frequency weighting factor for the record are set as “Frequency”. Variables used as a unique record identifier are set as “RecordID”.

The roles for A1\_Score, A2\_Score, A3\_Score, A4\_Score, A5\_Score, A6\_Score, A7\_Score, A8\_Score, A9\_Score, A10\_Score, Age, Gender, Juandice and PDD are set as Input as these are the independent variables. The role for Class/ASD is set as Target as this is the dependent variable. None of the roles are set as Both, None, Partition, Split, Frequency or RecordID.

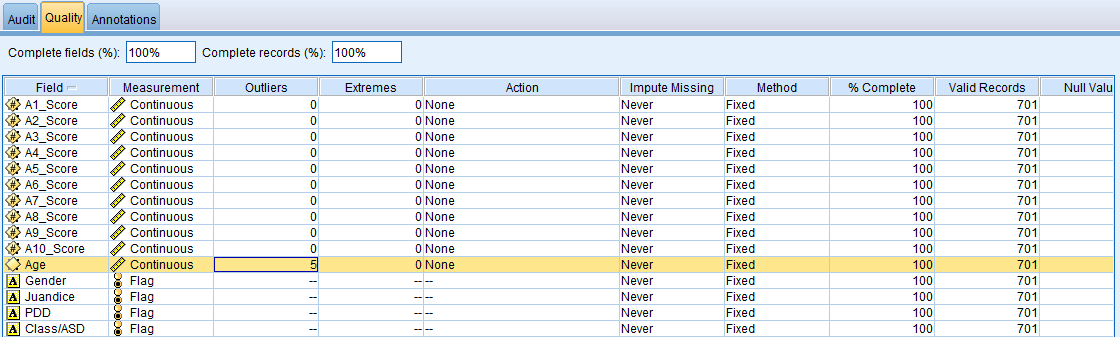
This is a classification problem as the Target is measured as Flag and the target variable to be predicted is non-metric, as opposed to an estimation problem, where the target variable would be metric or continuous instead.

**b)**



*Figure 2: Dataset characteristics for behavioural test scores*

As seen in Figure 2, the data range is consistent across all 10 fields of behavioral test scores, ranging from 0 to 1. The mean scores and standard deviation vary from 0.433 to 0.564 and 0.240 to 0.263 respectively, indicating that there are no particular issues as the scores are normally distributed, with no strong skewness to either side. The other fields of Gender, Jaundice, PDD and Class/ASD also appear to have no issues as the number of unique values is 2, reflecting the flag data measurement, where there can only be 2 distinct values. Hence, there are no obvious data entry errors since all data is within set parameters and of reasonable range.



*Figure 3: Data quality check for all fields*

The only exception appears in the Age field, as seen in Figure 3, where there are 5 outliers. Age ranges from 17 to 64, with a standard deviation of 9.712 and skewness of 1.035, indicating positive skewness whereby more subjects are younger in age. According to the default settings in Modeler, outliers are defined as being 3 standard deviations from the mean, while extreme values are 5 standard deviations from mean. As the mean age is 29.194, the 5 outliers indicate that there are 5 subjects aged 59 and above. However, these outliers do not have to be remedied as they are not unnatural and more importantly, are irrelevant inputs for testing of autism, which is generally agreed to be a condition that one is born with rather than developed over years.

Thus, the age of the subject taking the screening test is not of particular significance.

Also from Figure 3, records are shown to be 100% complete with 0 null values and all 701 records are valid for all fields. Hence, the data quality appears to be excellent with no issues such as missing data or data entry errors that could potentially cause problems in analysis.

**c)**

The objective of this analytics task is primarily to develop an ASD screening method that is not only effective, but also time-efficient and accessible. As it would be impractical to conduct a formal clinical diagnosis for every adult due to the costs and time required for the procedures involved, the industry needs a simpler method of distinguishing cases where further tests are necessary, but still with relatively high accuracy. Hence, this would be a predictive analytics task as the goal is to predict whether a subject exhibits traits of autism based on simple predictive variables or inputs such as scores on behavioural tests, family history and if they were born with jaundice.

Predictive analytics differs from descriptive analytics as the latter basically describes only past data i.e. what has already happened. Descriptive analytics is useful in summarizing large amounts of data and condensing what would otherwise be difficult to understand into a form that is easily interpretable. With descriptive analytics, one can understand the general characteristics or underlying patterns of the data at hand. Methods such as data visualization or association and clustering would fall under this category.

In the context of this problem, descriptive analytics can be conducted through clustering in order to group the data of past subjects into clusters. By grouping the data according to their degree of similarity, one would be able to identify the characteristics behind these groups and also whether they exhibit traits of autism. For instance, clustering may reveal 2 groups whereby the subjects in one group scored within a similar range across all behavioural tests and the subjects in the other group scored within a different range. Thus, health professionals would be able to conclude that a subject with a certain range of scores is more likely to display traits of autism than a subject without. Descriptive analytics could also simply show for instance that 60% of the subjects with a family history display traits of autism.

However, while descriptive analytics could provide valuable insights to aid future decisions, it is still limited to past data and cannot predict the likelihood of a particular subject displaying traits of autism. For this, predictive analytics is required, whereby past data is used to develop a predictive model that can forecast what would likely happen. By capturing relationships in a dataset, predictive analytics would allow health professionals to predict that a subject with a specific set of behavioural scores is highly likely to have traits of autism and should pursue further tests whereas another subject almost certainly does not based on his scores and the cost of additional tests would not be worth it.

The other type of analytics is prescriptive analytics, which is a step beyond predictive analytics. Not only does it predict a particular outcome, as in the case of predictive analytics, it also predicts multiple consequences based on a particular action and recommends the best course of action to take. Hence, prescriptive analytics can inform us of not just the “what”, but also the “when” and the “why”. Furthermore, prescriptive analytics can make use of both historical data and also real-time data feeds to continually improve prediction accuracy and suggest better decisions. This is a huge advantage over predictive analytics that has a model built on historical data and has to be updated regularly to account for environmental changes.

In the above example, prescriptive analytics would be able to optimize the entire screening process and maximize efficiency, cost-savings and also effectiveness of future treatments with regards to diagnosed patients. For instance, while predictive analytics may be able to identify the subjects that require further diagnosis, prescriptive analysis would be able to suggest the best order and timing in which they should go for testing. This could be dependent on the analysis on the subject’s profile whereby one case is more urgent, or it could be based on data input from the hospitals whereby appointments are redirected if a particular hospital is short on doctors at a particular point in time. Based on the subject’s test scores, the system may also be able to identify a specific doctor that has experience in dealing with the identified profile.

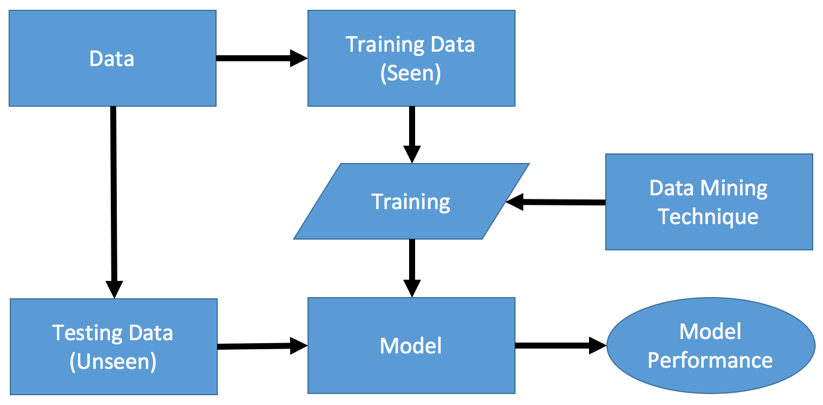
While prescriptive analytics would be more useful, it is also much more complex and requires significantly more data input. Therefore, as the objective is simply to predict whether a particular subject should pursue further tests for ASD based on available historical data of past subjects, it would be a predictive analytics task.

**d)**



*Figure 4: Partitioned training and testing sets*

Following the ratio of 80% and 20% for training and testing, the number of screening records in the resultant training and testing sets are 552 and 149 respectively, as shown in Figure 4 above. This ratio is appropriate for the given dataset as it can be considered to be sufficiently large at 701 records. The high training partition size of 80% will improve the predictive accuracy of the model whereas even at 20%, the testing set still has 149 “unseen” records, which should be sufficient for the model to test against.



*Figure 5: Training and testing framework diagram*

In the training and testing framework, a given dataset is partitioned into a training dataset and a testing dataset. The model is constructed using the training dataset, using data mining technique with defined parameters. The training data is considered “seen” by the model since it is used to construct the model. On the other hand, the testing data is considered “unseen” by the model given that it is not used to construct the model. The testing set will be used to check the validity of the model by verifying if it is able to give the correct prediction.

In predictive modelling, k-fold cross validation allows users to partition the sample into k equal subsets. 1 subset will be used as testing data for the model, whereas the remaining k - 1 subset will be used as training data for the model. The cross-validation is repeated k times, where each subset will be used for testing the model. This allows all of the observations to be used for training or testing at any given time. An average of the k performance metrics obtained from the cross validation pocess then serves as a final estimate of the model performance.

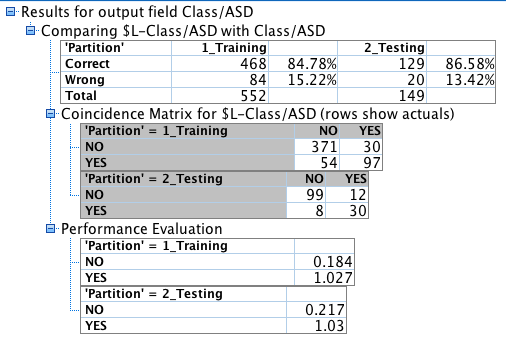
The k-fold cross validation should be used when the amount of available data is limited. A small dataset partitioned into training set and testing set will result in the testing set to be very small, which in turn will cause the predictions to have higher variance compared to a large testing set. By using k-fold cross validation, the problem with a small testing set is eliminated, since all the data in the dataset are used as testing set once, which allows the variance of the predictions to be lowered by averaging the results from the k subsets used as testing set.

The results from a k-fold cross-validation evaluation process can be interpreted by averaging the k performance results (which include measures such as accuracy rate, hit rate, sensitivity, etc) to produce a single estimation. Since the observations are partitioned into k equal subset and used as testing set for the k-fold cross validation, by averaging it, we will be able to obtain a single result than multiple results. The higher k is increased, the lower the variance of the performance estimate. Similarly, the lower k is decreased, the higher the variance of the estimate.

**e)**

In this case, the purpose of the logistic regression model is to explain the relationship of between one dependent binary variable and a few independent variables that are either categorical or continuous in nature. A binary dependent variable is defined as a variable with only two possible outcomes. Similarly, the target variable in this case study, ASD, has only two possible outcomes: “YES” or “NO”. In other words, the target variable, ASD, is dichotomous in nature. This fulfils the assumption that the target variable of the logistic regression model ought to have two distinct values only.

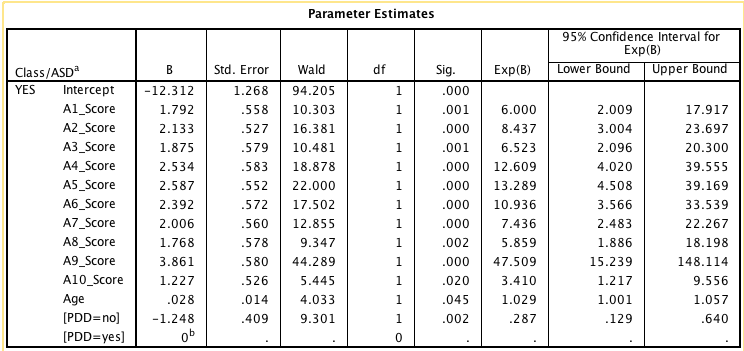
In this case, the analytics team is seeking ways to speed up the screening time and to improve the sensitivity, specificity and accuracy of the diagnosis process. Here, an understanding of the probability of a subject having the traits of ASD will accomplish that mission greatly. This is because quantitative results, especially in the area of human behaviour, can give the staff a better judgement of the state of the subject. For this reason, as the logistic regression model is designed to calculate the probability of occurrence of an event, the logistic regression model is appropriate for this case study.



*Figure 6: Logistic regression training and testing accuracy rate and coincidence matrix*

The overall training accuracy rate is 84.78% while the overall testing accuracy rate is 86.58%. The testing accuracy rate is generally high, which means that the model is acceptable. This is more impressive given the fact that the partition size is 80:20 because a smaller testing partition size may suggest that the model more accurate in its predictions. Thus, the model is generally acceptable in terms of its ability to predict the outcomes correctly.

The hit rate for event is 78.95% which implies that the accuracy of the model may actually be lower than what its testing accuracy rate suggests. However, on the other hand, the hit rate for non-event is higher at 89.19%, which may prove otherwise. Meanwhile, the sensitivity is revealed to be 71.43% which is significantly lower than the what the testing accuracy rate suggests. This implies that the logistic regression model may not be a good model in predicting the event (“YES”) and non-event (“NO”).



*Figure 7: Parameter Estimates*

The equation is Outcome of whether subject has ASD = 1.792 A1\_Score + 2.133 A2\_Score + 1.875 A3\_Score + 2.534 A4\_Score + 2.587 A5\_Score + 2.392 A6\_Score + 2.006 A7\_Score + 1.768 A8\_Score + 3.861 A9\_Score + 1.227 A10\_Score + 0.02815 Age - 1.248 [PDD=no] - 12.312. The logistics regression results of the “Forwards” method reveal that the p-values of most attributes range from 0.000 to 0.002. These relatively small p-values illustrate that the outcome of ASD is strongly determined by the degree of the inputs.

In this case, the strongest determinant of whether the subject has ASD is A9\_Score, which is the subject’s assessment score on the 9th carefully curated behavioural test. The A9\_Score attribute has the highest value of 3.861, which implies an increase in the probability of the subject having ASD traits by 47.509 times for every increase of 0.100 in the score in that test. Meanwhile, the weakest component of this model is revealed to be the age -- a logical discovery as the nature of ASD is understood to affect people of all ages. Nevertheless, the model illustrates that of all the attributes added in the “Forwards” method, the age attribute has the lowest value of 0.028.

Furthermore, the activity log in *Figure 1.2* indicated that the A9\_Score was the first variable to be added. As a rule in the forward method, the most significant input variable will be added first -- the extent of the variable’s significance depends on its R-squared value and how much the addition of the variable will increase the R-squared value. Meanwhile, the Age attribute was added in the last step, which signifies that it has the least effect on the outcome.

Another notable discovery is the [PDD=No] attribute. The negative value it possesses suggests that an inverse relationship between the [PDD=No] attribute and the likelihood that the subject has traits of ASD. In other words, it is less likely that the subject has traits of ASD if the family member does not have a pervasive development disorder.

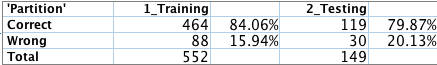
Not all the attributes are included in the equation because some of them like the “Jaundice” and “Gender” attributes were deemed to be too insignificant to determine the outcome of the subject’s ASD. They are reasonable omissions because these attributes are logically known and understood to have no causal relationship with whether the people will have ASD.

**f)**

Both the Logistic Regression model and the CHAID decision tree model engage in a stepwise approach to build their prediction models. In these models, the inputs are identified and are given a significance value based on the extent of effect each input has on the outcome. Upon comparison of the inputs’ significance values, the models will select the most significant input for further use. At this point, the two models differ in their use of the most significant input that they have selected.

The CHAID model comprises of the multi-way splits. The data gets partitioned according to the selected input and its split point. After that, the resultant split regions are re-analysed independently to produce further splits -- the process is continuous until all the inputs that are considered to be significant are chosen. On the other hand, the Logistic Regression model involves selecting an input (either the input with the highest or zero R square value or affects the equation’s R square value the most or the least, depending on the method used). Then it will either add or remove (or in stepwise selection, both) the selected input into the equation, one at a time. The process repeats itself until the inputs that are considered to be significant enough to affect the outcome are added or left in the equation.

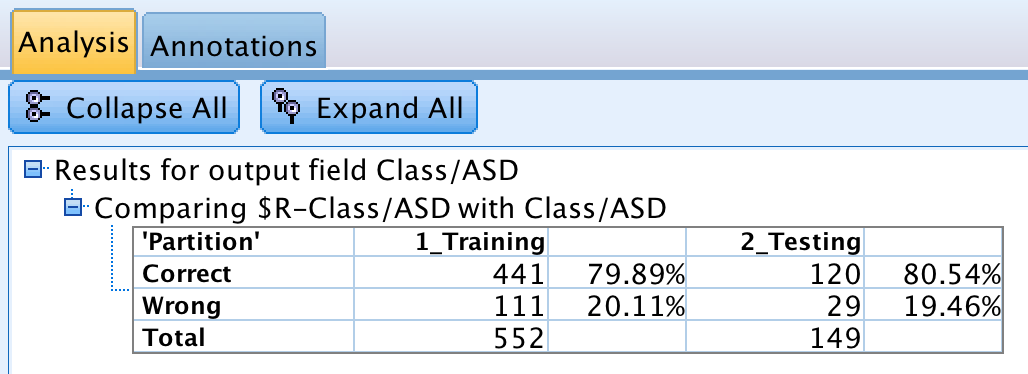
Furthermore, the CHAID tree is understood to be a classifier model while Logistic Regression model is seen to be more of a direct probability model than a classifier model. This is because the CHAID tree breaks down a dataset into smaller datasets based on the significant split point it has identified, which are then further broken down into even smaller datasets, hence forming classification groups. This is unlike that of the Logistic Regression model, where it instead generates probabilities of how likely the outcome is to occur based on all the significant inputs that it has identified. Lastly, the problem of multicollinearity seems to affect the accuracy of the logistic regression model more than the CHAID model. This is because of the CHAID’s ability to “drop” one column at each split, which the logistic regression model is unable to.



*Figure 8: CHAID decision tree training and testing accuracy rates*

The training accuracy rate is 84.06% while the testing accuracy rate is 79.87%. Both of CHAID model’s accuracy rates are lower than those of the Logistic Regression model (84.78% and 86.58% respectively). This suggests that the CHAID model is inferior to the Logistic Regression model in terms of its reliability to generate accurate predictions of the outcomes when provided with the relevant input.

**g)**

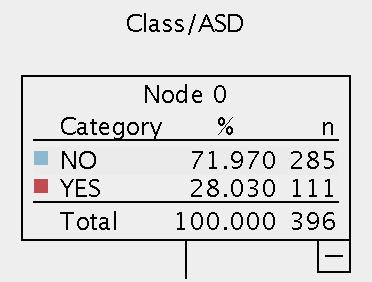


*Figure 9: C&RT decision tree training and testing accuracy rates*

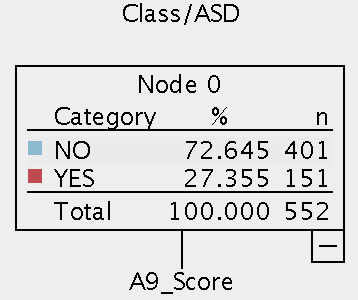
The training accuracy and testing accuracy of C&RT model is 79.89% and 80.54% respectively. Compared to the training accuracy and testing accuracy of Logistic Regression model of 84.78% and 86.58% respectively, the C&RT model did not perform as well as Logistic Regression model since both of its accuracy rates are lower. On the other hand, compared to the training accuracy and testing accuracy of CHAID model of 84.06% and 79.87% respectively, the C&RT model did not perform as well as CHAID model for training accuracy, but it performed better than CHAID model for testing accuracy.

The C&RT model may not be as effective as the Logistic Regression model and CHAID model for this problem. The C&RT model is a binary decision tree model where it only has two outgoing branches. It uses pruning which allows the C&RT model to prune nodes that do not contribute significantly to the prediction. As a result, when using C&RT model on this problem, only 2 inputs, A9\_Score and A6\_Score, are used when making the prediction. This is significantly fewer than both Logistic Regression model and CHAID model. Logistic Regression model uses 12 inputs when making the prediction, which are the assessment scores of all 10 behavioural tests, Age and PDD. CHAID model uses 9 inputs when making the prediction, which are the assessment scores of 7 behavioural tests, Juandice and PDD.

Although it is stated that both behavioural test scores and demographics characteristics have been shown to be effective in detecting ASD cases, the C&RT model only uses behavioural test scores but not demographic characteristics when making the prediction. This may cause the C&RT model to be less effective in detecting ASD cases since it does not take into consideration the demographic characteristics. Furthermore, it only uses 2 out of 10 of the behavioural test scores, which may cause it to be less accurate in making the prediction. On the other hand, both Logistic Regression model and CHAID model uses demographic characteristics when making the prediction. This allows it to be more effective in detecting ASD cases. Furthermore, the Logistic Regression model uses 10 out of 10 of the behavioural test scores and CHAID model uses 7 out of 10 of the behavioural test scores, which may allow it to be more accurate in making the prediction.



*Figure 10: C&RT node 0*



*Figure 11: CHAID node 0*

The total value of “n” in Node 0 of the CHAID model is 552, whereas the total value of “n in Node 0 of the C&RT model is 396. This total value of “n” represents the total number of observations used in the training data for the chosen model. The reason for the difference in the total value of “n” in Node 0 between CHAID model and C&RT model is due to the C&RT model having a 30% Overfit prevention set parameter, which places a limit on the number of observations to be used. The CHAID model does not have a 30% Overfit prevention set parameter, which allows it to use the maximum number of observations.

**h)**

Siemens is a German conglomerate and leader in industrial manufacturing, with several divisions under it such as “Power and Gas” which deals primarily with the energy industry. This means that the division’s main product portfolio consists of a variety of gas and steam turbines, to be sold to customers for various purposes such as for power plants, industrial power generation and offshore oil production etc.

In recent years, the lower cost and increased viability of renewable energy sources have led to structural weaknesses as demand for traditional energy sources and in turn gas and steam turbines has declined drastically. Similar to its competitors such as G.E., Siemens’ Power and Gas division has engaged in cost cutting by reducing headcount while streamlining processes to maximize profits from a now smaller market.

Hence, Siemens faces the business problem of having to win as many bids and opportunities as possible, but with significantly less resources. It is thus crucial that the limited resources are dedicated only to opportunities that Siemens stand a reasonable chance of winning as costs of engaging clients and proposal preparation would be wasted otherwise. Predictive modelling can help to identify these opportunities by forecasting the likelihood of Siemens winning a particular opportunity, given the predictive variables.

Siemens could construct a predictive model built on historical data of opportunities pursued over the past 5 years, which would number in the high hundreds considering that it operates in most regions in the world. The format of the dataset to be used would simply be a .csv file as data is always recorded in excel and the software Salesforce can also export relevant data in this format. Possible inputs to utilize would be both non-metric ones such as the country the project was located in, the particular model of turbine proposed, whether the bid was made together with a consortium, as well as metric ones such as the value of the potential contract or forecast gross margins. While there are several stages of negotiation involved, for simplicity’s sake, the target output can be classified as either opportunity won or opportunity lost.

For this predictive modelling, an appropriate method would be the CHAID decision tree since it can analyse both metric and non-metric inputs. Furthermore, not only is it able to deal with a large dataset, it also allows for multiway splits, which would be necessary for data and inputs as complex as this. For instance, the turbines would not only be split into nodes of either gas or steam, but also into more specific ones such as the range of megawatts it can generate or the specific frequency required by the customer.

In terms of possible results, the predictive model could shed light on neglected turbine models or countries without any strong customer connections but with surprisingly high win rates. Thus, by using this predictive model on the opportunity pipeline, Siemens’ sales team would be able to better choose opportunities to pursue, and minimize costs wasted on opportunities unlikely to be won.

**References**

Olson, D. L., & Shi, Y. (2007). *Introduction to business data mining*. London: McGraw Hill.

**Appendix**

|  |  |  |
| --- | --- | --- |
| **Testing scenario** | Hit Rate | Accuracy Rate |
| Event | Hit Rate for Event  = 30/ (8 + 30)  = 0.789473  78.95% | Accuracy Rate for Event  = 30/ (30 + 12)  = 0.714285  71.43% |
| Non-event | Hit Rate for Non-event  = 99/ (99 + 12)  = 0.89189  89.19% | Accuracy Rate for Non-event  = 99/ (99+8)  = 0.9252336  92.52% |