Paper Number:217-2010

Using the SAS® System and SAS® Enterprise Miner™ for Data Mining: A study of Cancer Survival at Mayo Clinic

Leonard Gordon, University of Kentucky College of Public Health, Lexington, Kentucky

ABSTRACT

This paper evaluates and predicts a certain epidemiological (cancer survival) condition using data-mining techniques in SAS®. A data set that contains information about the survival of lung-cancer patients from a study at the Mayo Clinic was extracted from the R survival package. Data-mining techniques—namely linear and logistic regression models, regression and classification trees, and nearest-neighbor analysis—are used for the analysis to see which method is best for determining cancer survival. Both a continuous response variable and a dichotomous response variable are selected and used to evaluate cancer survival of the patients. Linear regression, regression trees, and nearest-neighbor analysis are used to analyze the continuous response variable; logistic regression, classification trees, and nearest-neighbor analysis are used for the dichotomous response variable. The results show that the dichotomous response variable using the nearest-neighbor analysis (using 8 neighbors) and the classification trees are best for the analysis of cancer survival in this particular data set, with a correct classification rate of 78.6% for both methods.

INTRODUCTION

The data set is on the survival of lung cancer of 228 patients from a study at the Mayo Clinic. It constitutes 10 variables, namely: inst, time, status, age, sex, ph.ecog, ph.karno, pat.karno, meal.cal and wt.losss. Inst is the institution code of where the patient was treated, time is the survival time in days, status is the censoring status, age is the age of the patient in years, sex is the gender of the patient, ph.ecog is the ECOG performance score, ph.karno is the Karnofsky performance score as measured by a physician, pat.karno is the same score as measured by the patient, meal.cal is the amount of calories consumed at meals and wt.loss is the amount of weight lost or gained in the last six months. The (censoring) status has values 1 and 2, where 1= censored and 2=dead. The gender of the individuals have values 1 and 2, where 1=male and 2=female. The ECOG (Eastern Cooperative Oncology Group) score is used by doctors and researchers to assess how a patients' disease is progressing, how the disease affects the daily living abilities of the patients and to determine appropriate treatment and prognosis. It has values ranging from 0=good to 5=dead. The Karnofsky performance scale index allows patients to be classified as to their functional impairment. It is used to compare effectiveness of different therapies and assess the prognosis in individual patients. The lower the Karnofsky score the worse the survival for serious illnesses. The variables time, age, ph.karno, pat.karno, meal.cal and wt.loss are continuous variables while status and sex are dichotomous variables. Inst and ph.ecog are neither continuous nor dichotomous, so for analysis purposes they are considered as ordinal. For analysis purposes, the '.' was taken out of the variable names that contained them, as SAS® had a problem reading the variable names. So the variables ph.ecog, ph.karno, pat.karno, meal.cal and wt.loss were converted to phecog, phkarno, patkarno, mealcal and wtloss respectively. The data set was extracted from the R survival package.

The continuous outcome in this data set is wtloss and the dichotomous outcome is status. Wtloss was chosen because weight is affected a lot by cancer. Usually, a patient would lose weight due to treatment but there are other instances in which they gain weight. Hence, this is an indication of the severity of the disease in the individual. Status is already an indication of the extent of the disease in the patient, as it tells us whether the patient is alive or dead. So it is regarded as a fitting response variable.

A good rational for the data mining with regards to this data set is that we can be able to possibly prevent death in cancer patients who are still alive which is an inevitable result in most cases of cancer if the disease is left unattended. As a result, we can be able to determine the extent of cancer in the patients with regards to the available explanatory variables and try to prevent the inevitable and prolong the life of the patients. Furthermore, new ways of data mining are explored and compared with already existing methods.

EXPLORATORY DATA ANALYSIS

The UNIVAR macro, one of Fernandez's macros, was applied to the data and some exploratory analysis was done. Most of the continuous variables set are slightly skewed to the right but are assumed normal for the purpose of analysis. Cancer survival (time) is right skewed as different patients can survive a wide range of days depending on the stage and severity of their disease. Age was normally distributed with most of the patients being middle-aged. There were two suggestions for outliers, but these were not deleted from the data due to the fact that cancer is not a respecter of age, even infants suffer from cancer. Mealcal was also found to be right skewed as different patients have different appetites and eating habits. On average, it is advised that an adult should take about 3000 calories a day which would translate to about 1000 calories per meal. So the top five values for mealcal were considered outliers as they were flagged by the macro. This resulted in five patients being deleted from the data set. Wtloss is also right skewed as different patients can lose or gain weight depending on their metabolism and the effects of the disease on their anatomy. However, since all the patients were middle-aged or higher, some of the weight loss numbers tend to raise an eyebrow as it is commonly acceptable knowledge that as one ages it is more difficult to lose weight as the metabolism tends to slow down. On the other hand, patients usually lose weight during cancer treatments due to the effect of cancer on the body and loss of appetite. But, it was discovered that there are instances where they also gain weight due to medication and hormone therapy. As a result, there were no outliers deleted due to weight even though the macros suggested some values.

In total, 5 patients were considered outliers and deleted from the data set which reduced the total number of observations to 223.

After the data set had been “cleaned” the RANSPLIT macros was applied to divide it into training, validation and test data. The training, validation and test data had 111, 56 and 56 observations respectively. The training subset was used to fit various competing models, the validation subset was used to judge between and choose from among the competing models. Then the test subset was then used to evaluate the chosen model.

PREDICTION OF THE CONTINUOUS OUTCOME

The three methods that were used to evaluate the continuous outcome (wtloss) are linear regression, regression trees and nearest neighbor analysis.

i. Linear Regression

The REGDIAG macro was applied to the training data to perform multiple linear regressions. It was observed that there was no correlation between wtloss and time. This was very surprising as one would think that there is some degree of association between the two variables. However, a possible explanation for that could be the fact that there are a lot of missing variables in the data set. Another possibility could be that time and weight are not linearly related, so a transformation of one or the other in a model could be better. Also, a strong correlation between patkarno and phkarno had been anticipated as they both measured the same thing, in which case the number of explanatory variables in the model would have been reduced. This was not the case. The correlation between patkarno and phkarno was not really strong. So both variables were left in the model.

The model with all the explanatory variables was not such a good fit to the data so three other models were chosen based on C(p), AIC (Akaike Information Criterion) and SBC (Schwarz-Bayesian Information Criterion) and these were labeled Models I, II and III respectively. Model I had a C(p) value of 2.7511, Model II an AIC value of 379.2632 and Model III an SBC value of 389.3903. The models are as follows:

Model I: wtloss =20.495 – 6.557sex + 4.0368phecog – 0.006mealcal
Model II: wtloss= 22.288 – 0.266inst – 6.293sex + 4.070phecog – 0.006mealcal
Model III: wtloss= 5.869 + 4.451phecog

Based on the assessment of R^2 values on the validation data set, model II was chosen as the best model.

Due to the very small R^2 values for the model with all the explanatory variables a transformation was considered to see if the situation could be improved. As suggested by the augmented partial residual plots from the REGDIAG macro, the quadratic terms of some of the explanatory variables were added to the original multiple linear regression model. This resulted in the quadratic terms for age, time, phkarno and mealcal being added to the original model and these were labeled agsq, tmsq, pksq and mcalsq respectively. This only increased the R^2 value for the model with all the explanatory variables slightly. While this was better than the previous R^2, it is known that an increase in the number of explanatory variables causes an increase in the R^2 value. Added to that, due to the complexity of interpreting transformed variables the models were left as is.

ii. Regression Tree

Regression trees have an added advantage over linear regression models in that they can be employed when the relationship between the expected response and the explanatory variables are not linear without having to worry about the transformations of the explanatory variables. Additionally, they take into consideration interactions between the explanatory variables.

Using SAS® Enterprise Miner™ a regression tree was used to evaluate wtloss as an outcome given the other explanatory variables. In predicting the wtloss of a patient only the explanatory variables mealcal, sex, phkarno, age, phecog and patkarno are used. It is interesting to note that time does not play an important role in predicting wtloss. The questions are chosen so that the average squared error on the training data is minimized. Further analysis indicates that the regression tree is not a very good fit for the model.

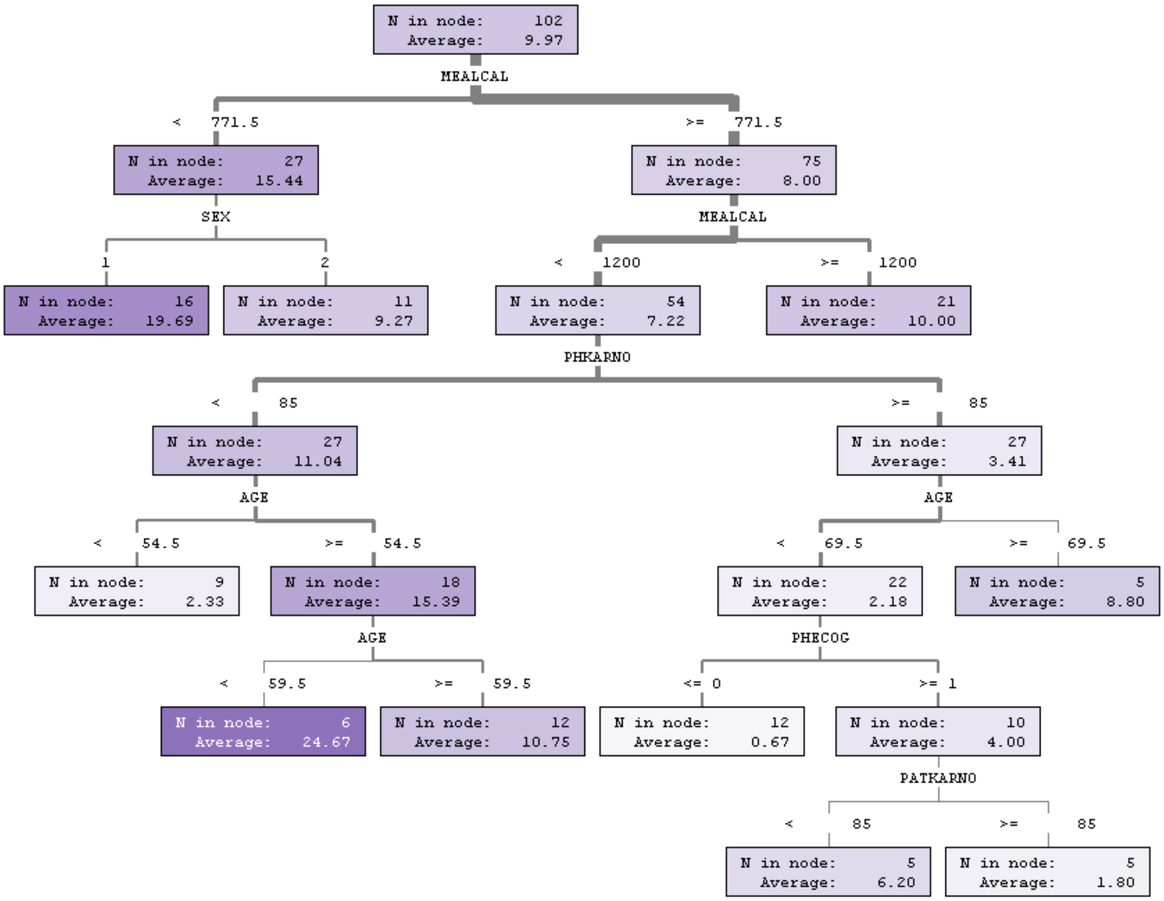

Figure 1 – Regression Tree for prediction of continuous outcome

Although the regression tree is not a good fit for this particular model, it could be employed in other data mining and predictive procedures as it is less quantitatively involved. There is no need for quantitative models and the nodes can be easily followed to make predictions.

iii. Nearest Neighbor Analysis

Nearest neighbor analysis entails indentifying one or more subjects in the data set who are similar on the explanatory variables to the subject for whom a prediction is desired. These similar subjects are called nearest neighbors, hence the name nearest neighbor analysis. When there are multiple nearest neighbors, prediction for the new subject is based on the average value of the response variable among the neighbors. Similarity is calculated by a similarity score defined by one divided by the sum of one and the squared difference between the new subject and the neighbor for all the explanatory variables involved.

However, there are some difficulties in defining similarity which include scaling, relevance and dimensionality. Depending on how the variables are scaled, this can change the similarity score considerably. For example, if age is measured in decades as opposed to years certain subjects can be found to be more similar to the new subject. Also, it is hard to determine which of the variables are relevant to the response variable. A source of contention is whether all the explanatory variables should be included in the calculation of the similarity score. Finally, there is a problem with dimensionality in that if the number of explanatory variables is large we can have high similarity scores for subjects who are quite different from the new subject for whom we want to make a prediction.

The number of nearest neighbors can be specified or determined. To determine the number of nearest neighbors perform nearest neighbor analysis several times with different values of k and then choose the value of k for which the average squared error is minimized.

Nearest neighbor analysis was applied to the regression problem using SAS® Enterprise Miner™. 8, 16, 20 and 32 neighbors were tried and the respective average squared errors on the validation data set were 186.30, 176.45, 181.18 and 190.03 respectively. The reason 20 neighbors was tried was to see if an increase from 16 neighbors would produce a better result, but that was not the case. The best performance on the validation data set was given by the 16 neighbors' analysis as indicated by the smallest average squared errors.

For the prediction of the continuous outcome (wtloss) the best method was found to be the 16 neighbors nearest neighbor analysis as it produced the highest R^2 value on the validation data set amongst the competing models. Even though it is not that much better than the other models, it proved to be the best.

PREDICTION OF THE DICHOTOMOUS OUTCOME

The three methods that were used to evaluate the dichotomous outcome (status) are logistic regression, classification trees and nearest neighbor analysis.

i. Logistic Regression

The LOGISTIC macro was applied to the training data set to perform logistic regression with status as the outcome. Logistic regression was done on the model with all explanatory variables and two other models were chosen based on AIC and SBC. These models were labeled models I II and III respectively. The models are as follows:

Model I : log(p/(1-p))=1.1407 - 0.0242*inst - 0.00182*time + 0.0270*age -0.3266*sex + 1.0527*phecog + 0.0257*phkarno – 0.0356*patkarno – 0.00026*mealcal – 0.0213*wtloss
Model II : log(p/(1-p))=4.9759-0.0428*patkarno – 0.00149*time
Model III: log(p/(1-p))= 4.9268 – 0.0479*patkarno

Where p = the probability (risk) of status = 2.

Based on the validation data set, the best model chosen was model I as it had the highest correct classification rate of 75.6% compared to the other two models.

ii. Classification tree

Classification trees have an added advantage over logistic regression models in that they can accommodate a response variable with more than two categories. They are also better in that they can be employed when the relationship between the explanatory variables and the logit risk is not linear and there may be interactions between the explanatory variables.

Using SAS® Enterprise Miner™ a classification tree was used to evaluate status as an outcome given the other explanatory variables. The classification tree only uses time and inst to make a prediction on the status of a patient. It is interesting to note that the longer one lives they have a lesser chance of being dead due to the cancer. This might be as a result of the advancement in cancer treatment that is widely available these days. Also, depending on the institution that one finds themselves there is a higher probability of being dead from cancer. This might add to the hypothesis of the medical disparities among different health institutions based on geography and socio-economic status. The correct classification rates on the validation and test data sets are 78.6%.

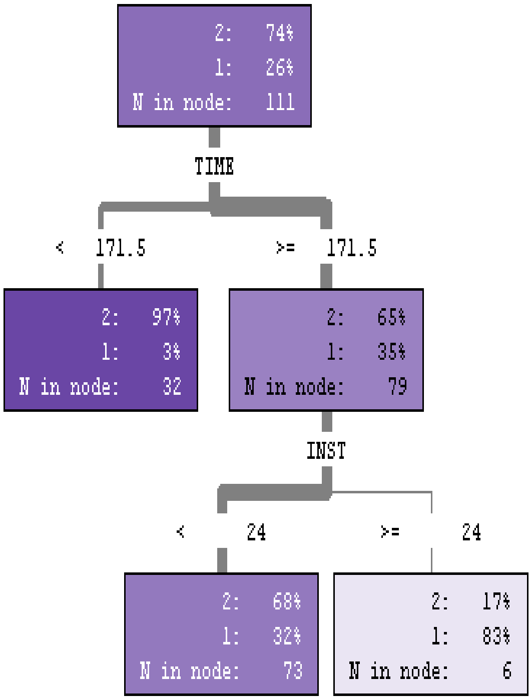

Figure 2 - Classification tree for the prediction of the dichotomous outcome

iii. Nearest Neighbor Analysis

Nearest neighbor analysis for the classification problem is similar to the scenario described above in the prediction of the continuous outcome. The difference being that prediction is now made based on the category of the response variable that appears most frequently among the k nearest neighbors.
Nearest neighbor analysis with 8, 16, 20 and 32 neighbors was applied to the data set using SAS® Enterprise Miner™. The best performance on the validation data set was given by the 8 nearest neighbor analysis as it had the smallest misclassification rate on the validation dataset. Hence, the correct classification rate for the 8 nearest neighbor analysis is 78.6%.

For the prediction of the dichotomous variable (status) there was a tie between the classification tree and the 8 nearest neighbor analysis for the correct classification rates on the validation data set. Both attained a correct classification rate of 78.6%.

DISCUSSION AND CONCLUSION

One of the main problems with the data mining was that there were a lot of missing variables. For example, when the REGDIAG macro was applied to the training data set out of the 111 observations only 78 could be used. If it had been in my power and time would have allowed, the missing values would have been updated so that there would be a complete clean data set. Also, it would be great to have a data set with a larger sample size so that when the data is divided in training, validation and test data sets there would be more observations.

Another possible problem with the data set was that the variables were not all linearly related. For example, it was very puzzling to find out that there was no correlation between time and weight. This could have probably been rectified by applying different transformations, other than the quadratic terms, to the variables and observing how the models performed given the transformations. A possible problem with that would be the interpretation of the transformations. However, given time and a lot of available resources this could be overcome.

For further consideration of the data, the idea that the institution where one is treated affects ones medical outcome as suggested by the classification tree could be looked and assessed for statistical significance.

On the whole, the prediction of the dichotomous variables turned out to be more successful than the prediction of the continuous variable. So in the event of trying to predict cancer based on this data set, the dichotomous variable would be used.

Furthermore, some alternative methodologies were considered to gain a complementary perspective in the prediction of the outcome variables. Regression and classification trees and nearest neighbor analysis were used to compliment the already established linear and logistic regression models. Some of these methodologies were less quantitatively involved and can be employed as more user friendly approaches for statistics for the common man.

REFERENCES

Fernandez, George. Data Mining Using SAS Applications. Chapman & Hall 2003.

Macros. www.agr.unr.edu/gf/dm.html

Eastern Cooperative Oncology Group. http://ecog.dfci.harvard.edu/general/perf_stat.html. 1998-2000.

Karnofsky Performance Status Scale Definitions Rating (%) Criteria. http://www.hospicepatients.org/karnofsky.html

National Cancer Institute, US National Institutes of Health. .http://www.cancer.gov/cancertopics/eatinghints/page3

ACKNOWLEGDEMENTS

I would like to thank Dr. Charnigo of the College of Public Health University of Kentucky for his support in writing this paper and the College of Public Health for their sponsorship.

CONTACT INFORMATION

Name: Leonard Gordon
Enterprise: College of Public Health, University of Kentucky, Lexington, KY
Address: 121 Washington Ave., Ste 203A
City, State ZIP: Lexington KY 40536
Work Phone:
Fax: 859-257-6644
E-mail: leonard.gordon@uky.edu