Ryan Kuhn

Special Topics in Audit Analytics

Project Proposal

As part of the Special Topics in Audit Analytics course I am required to complete a course project demonstrating the tools and techniques learned during the course of the semester. I am specifically interested in learning to use apriori algorithm and association rules. This analytical method is new to me and it looks to be a very powerful tool. I will also like to incorporate other models such as regression and classification to have a complete analysis.

For my dataset, I intend to use the Credit Approval Dataset from the UCI Machine Learning Repository. The dataset is a mix of continuous and categorical data which have been converted to meaningless variables by the data curators to protect the privacy of the credit applicants. There are over 600 instances and 16 observations for each instance. Approximately 5% of the observations are missing values which will need to be addressed. I think the dataset is appropriate for this project because it lends itself to a diverse set of models. The continuous variables are well suited for regression analysis; the categorical variables are suited for association rules; together the data can be modeled with a classifier algorithm. The dataset can be found here under ‘Credit Approval’: <http://archive.ics.uci.edu/ml/datasets.html?format=&task=&att=&area=bus&numAtt=&numIns=&type=&sort=nameUp&view=table>

My research questions will be a bit fictional because the dataset was converted to meaningless values. I will have to assume, for example, that in a credit approval dataset that one of the categorical values will be a status question such as race or gender. I will also have to assume that one of the continuous variables is related to income levels.

1. Is there a correlation between the continuous (numerical) values and the credit approval status? Can this relationship be used to predict if a person is granted credit? If yes, does the relationship indicate reasonable risk management strategies?
2. If we assume one of the categorical variables indicates a protected status such as race, gender, or religion, then is there a statistically significant difference in how credit is granted that could indicate a compliance risk from bias or discrimination? Contrarily, could the difference indicate a business opportunity for the lender?

To accomplish the objectives of this analysis, I will be using R and its diverse set of packages available to perform an analysis. The first of which is the base R platform which includes the glm() function. GLM is short for generalized linear regression which I will use to fit the numerical values. The second tool is the apriori()function in the arules package. This function is the implementation of the apriori algorithm within R. The association rules will be useful for identifying patterns in the categorical values and can be used to fill in missing values. Finally, I will use the rpart() function in the rpart package, to creates a decision tree over the categorical items in thedataset. Throughout the analysis I will be using the ggplot2 package for visualizations. Once all three are completed, I will try to improve the performance by weighting and combining the models. Generally speaking, the combined performance of several good performing algorithms will be better than a single great algorithm.

In your final report, you may need theoretical support (e.g., prior study etc. ) to explain your findings.

-------Ting Sun